Summer 2015

Characterization and Decoding of Speech Representations From the Electrocorticogram

Shreya Chakrabarti
Old Dominion University

CHARACTERIZATION AND DECODING OF SPEECH REPRESENTATIONS FROM THE ELECTROCORTICOGRAM

by

Shreya Chakrabarti
B.Tech. May 2011, West Bengal University of Technology

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

ELECTRICAL AND COMPUTER ENGINEERING

OLD DOMINION UNIVERSITY
August 2015

Approved by:

Dean J. Krusienski (Director)

Shuiwang Ji (Member)

Jiang Li (Member)

Dimitrie C. Popescu (Member)

ABSTRACT

CHARACTERIZATION AND DECODING OF SPEECH REPRESENTATIONS FROM THE ELECTROCORTICOGRAM

Shreya Chakrabarti
Old Dominion University, 2015
Director: Dr. Dean J. Krusienski

Millions of people worldwide suffer from various neuromuscular disorders such as amyotrophic lateral sclerosis (ALS), brainstem stroke, muscular dystrophy, cerebral palsy, and others, which adversely affect the neural control of muscles or the muscles themselves. The patients who are the most severely affected lose all voluntary muscle control and are completely "locked-in," i.e., they are unable to communicate with the outside world in any manner. In the direction of developing neuro-rehabilitation techniques for these patients, several studies have used brain signals related to mental imagery and attention in order to control an external device, a technology known as a brain-computer interface (BCI). Some recent studies have also attempted to decode various aspects of spoken language, imagined language, or perceived speech directly from brain signals. In order to extend research in this direction, this dissertation aims to characterize and decode various speech representations popularly used in speech recognition systems directly from brain activity, specifically the electrocorticogram (ECoG). The speech representations studied in this dissertation range from simple features such as the speech power and the fundamental frequency (pitch), to complex representations such as the linear prediction coding and mel frequency cepstral coefficients. These decoded speech representations may eventually be used to enhance existing speech recognition systems or to reconstruct intended or imagined speech directly from brain activity. This research will ultimately pave the way for an ECoG-based neural speech prosthesis, which will offer a more natural communication channel for individuals who have lost the ability to speak normally.

ACKNOWLEDGMENTS

First of all, I would like to acknowledge the invaluable guidance, support and encouragement I received from my advisor, Dr. Dean J. Krusienski throughout the course of my dissertation. Without his resolute belief in my abilities and dedicated involvement, this research would not have been possible. I would like to thank Dr. Gerwin Schalk, Research Scientist at the Wadsworth Center for providing me with this unique data set and for his direction and help throughout this research. I am also indebted to Dr. Jonathan S. Brumberg, Assistant Professor of Speech-Language-Hearing at University of Kansas, for his active involvement, constant guidance, and perceptive insight throughout the course of this dissertation, in particular, in abstracting ideas and interpreting the outcomes of this study.

I am fortunate to be a part of the Advanced Signal Processing in Engineering and Neuroscience (ASPEN) Laboratory at Old Dominion University, whose members have all directly or indirectly helped me in making this dissertation a success. In particular, I would like to acknowledge the support and encouragement I received from Garett D. Johnson, Komalpreet Kaur, and Yalda Shahriari, who helped me stay motivated during my PhD. I am also very grateful to my dissertation committee members for their insightful discussions and suggestions, which helped me improve the quality of this dissertation.

I am also grateful to the invaluable love, kindness, and support I received from my friends during this time, which was a great source of strength for me throughout my PhD studies. Finally, I would like to acknowledge, with gratitude, my parents for their unconditional love, unwavering enthusiasm, and steadfast encouragement during this time, without which this dissertation would not have been possible.

The work described in the following dissertation was supported in part by the National Science Foundation (1064912, 1451028).

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Close to two million people in the United States, and far more all over the world, are affected by neuromuscular disorders such as amyotrophic lateral sclerosis (ALS), brainstem stroke, spinal cord injury, muscular dystrophy, multiple sclerosis, cerebral palsy, and others [1]. These disorders disrupt the neuromuscular pathways through which the brain communicates with and controls the muscles in the rest of the human body [2]. The patients who are the most severely affected lose all voluntary muscle control, and are completely "locked-in", i.e., they are unable to communicate with the outside world in any way. Modern day life-support and rehabilitation technologies allow patients of neuromuscular disorders, including those who are completely locked-in, to continue living for many years. However, such systems add greatly to the personal, social and financial burdens of patients and caregivers.

Another possible mechanism can help restore functionality to these patients, to a certain extent, by providing the brain with an alternative non-muscular pathway to communicate with and control the external environment - a technology known as a brain-computer interface (BCI). A BCI system can help restore, enhance, supplement or replace the brain's natural communication with the outside world, in turn helping these patients lead a better life [2].

1.1 BRAIN-COMPUTER INTERFACES

A BCI is a communication system in which the commands that the brain sends to the external world do not go through the natural pathways of nerves and muscles; instead, they are decoded by an external system, which then translates the commands and uses them to control an appropriate device. There are four main components of a BCI system: Signal Acquisition, Feature Extraction, Feature Translation and Device Control (see Figure 1).

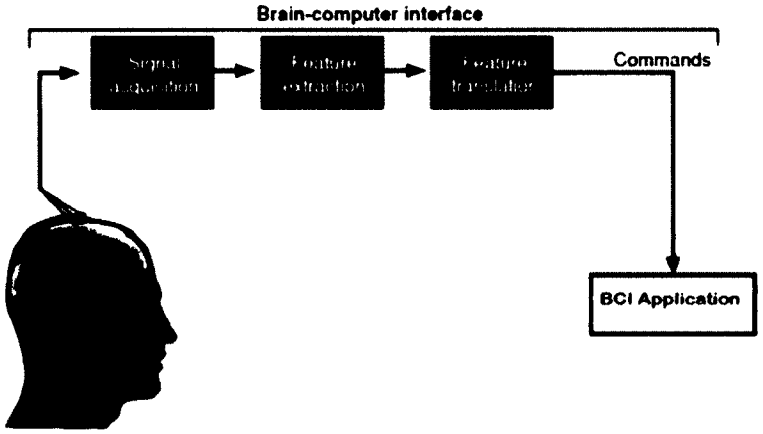


FIG. 1: An overview of the basic components of a BCI system. Electrical signals are acquired from the brain, after which they are analyzed to measure signal features that reflect the user's intent. These features are translated into commands that operate external application devices that replace or enhance natural central nervous system outputs. Based on figure in [3].

The neural activity of the user can be acquired using a variety of modalities, some of which are shown in Figure 2. These neural data acquisition modalities are broadly classified as invasive or non-invasive. Non-invasive methods, such as electroencephalography (EEG), functional magnetic resonance imaging (fMRI) and

functional near infrared spectroscopy (fNIR), record brain activity without penetrating the scalp and do not require surgery. However, these non-invasive techniques can suffer from issues such as comparatively low signal-to-noise ratio (SNR), spatial resolution, and/or temporal resolution. In contrast, invasive methods such as electrocorticography (ECoG) or stereotactic depth electrodes record electromagnetic potentials either from the surface of the cortex or from deeper layers within the brain, respectively. These techniques generally provide an amenable balance of SNR and spatial and temporal resolution compared to non-invasive techniques.

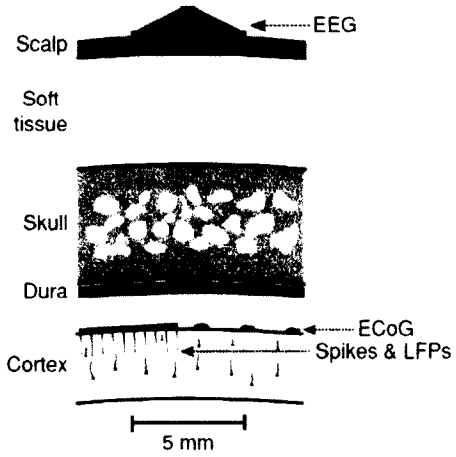


FIG. 2: Recording sites used by BCI systems. Adapted with permission from [3].

Once the neural signals have been acquired, they are analyzed and decoded in the feature extraction and translation modules. Usually, feature extraction is preceded by a pre-processing stage wherein artifacts that typically contaminate the acquired signals are removed in order to boost the SNR. Next, feature extraction is performed; this is the process of extracting certain characteristics from the acquired brain signals that reflect the user's intent. Following feature extraction, feature translation is

performed, which generally involves the application of a classification or modeling algorithm to predict the intention of the user. The results of this prediction are then translated into appropriate commands for the output device.

1.2 MOTIVATION

The primary aim of brain-computer interfaces is to help patients communicate with the external environment in a fast and intuitive manner [2]. Most BCI studies to date have focused on extracting intended movement, imagined movement or attentive selection in order to control a cursor [4], to type characters on a screen[5], or to control a prosthetic limb [6,7]. However, the existing fastest BCI for communication reports an information rate of 2.1 bits/ second [8], which is much lower than the average natural speech production rate of 25 bits/ second [9]. Hence, for a BCI to truly achieve real-time communication rates, it is desirable to have a BCI that approaches the rates of natural speech. Also, a BCI will only be truly intuitive if it can decode the meaning of, or the semantic information present in, the thoughts of the subject directly from cortical activity. A BCI which uses imagined speech in order to control external devices may thus be the gold standard for such an intuitive and practical BCI. In the direction of this ultimate aim, it is important to first understand the relationship between cortical activity and acoustic speech, the semantic information present in speech, and different representations of speech. After this, it would be useful to examine the possibility of predicting various components of articulated or imagined speech directly using brain signals [10]. This dissertation aims to address these two important research questions by exploring the use of ECoG to:

a. Characterize the spatio-temporal relationships between cortical activity and various speech representations popularly used in speech recognition systems as the brain prepares, produces and perceives self-produced, continuous speech.

b. Decode these speech representations directly from the cortical activity using simple and optimized feature sets and modeling techniques.

1.3 PRIMARY CONTRIBUTIONS OF THIS DISSERTATION

A majority of earlier studies used relatively simple components of speech such as vowels, phonemes, words, spectrograms, etc. for ECoG-based speech decoding. The first contribution of this dissertation is that it provides a comprehensive summary and review of earlier ECoG-based speech studies, as described in Chapter 2 [11]. An important contribution of this dissertation is that it characterizes and decodes speech representations commonly used in automatic speech recognition systems directly from ECoG activity. These representations are based on more fundamental acoustic features that can better facilitate generative models of speech from ECoG. Thus, this dissertation bridges the gap that currently exists between ECoG-based speech BCIs and speech recognition systems. This analysis is categorized based on the type of speech representation being studied as follows:

1. Characterization and decoding of production-based speech representations from ECoG activity: Various production-based speech representations such as the speech power envelope, the fundamental frequency (pitch), formants, and the linear prediction coefficients, which are all based on speech production models,

are characterized using spatial and temporal correlation metrics and decoded using ECoG activity.

2. Characterization and decoding of perception-based speech representations from ECoG activity: Various perception-based speech representations such as the mel frequency cepstral coefficients (MFCC) and the perceptual linear prediction (PLP) coefficients, which are both based on speech perception models, are characterized using spatial and temporal correlation metrics and decoded using ECoG activity.

These analyses uniquely highlight the spatiotemporal evolution as well as the common neural bases of speech in the brain at a high temporal resolution, which was previously not possible using other imaging modalities. Another unique aspect of this dissertation is the use of continuous speech for characterization and decoding, which differs from prior speech-based ECoG BCI research that evaluated discrete components such as words and phonemes. Continuous speech is more representative of natural daily communication; hence, it is more practical for creating generative models of speech toward the development of a natural speech prosthesis. This research also bridges the gap between speech articulation-based and speech perception-based ECoG studies by studying the underlying neurophysiology for both these conditions in terms of speech characterization and prediction. Finally, while the majority of earlier studies primarily employ wide-band gamma frequencies for decoding various speech components from ECoG activity, this work introduces more optimized feature sets derived from narrower gamma sub-bands to improve upon existing ECoG-based

speech decoding models. These collective findings provide important insights toward the development of a real-time speech prosthesis using ECoG.

1.4 DISSERTATION OUTLINE

The remainder of this dissertation is organized as follows. Chapter 2 discusses the background for this study, primarily highlighting the state of the art in current ECoG-based speech decoding studies, and discussing popularly used speech recognition techniques, since this dissertation attempts to mainly bridge the gap between these two research areas. Chapter 3 covers the dataset analyzed in this dissertation including the experimental paradigm, data acquisition, and data analysis. The characterization and decoding of production-based and perception-based speech representations from ECoG are detailed in Chapters 4 and 5, respectively. Following this basic characterization and development of decoding models, the optimized characterization and decoding models for two fundamental speech components, power and fundamental frequency, are presented in Chapter 6. Chapter 7 concludes the dissertation with a discussion of the main contributions and possible future directions of this research.

CHAPTER 2

BACKGROUND

This chapter provides a comprehensive and detailed review of existing ECoG-based speech characterization and decoding studies. A majority of the content in this chapter is derived from [11].

While the neural correlates of speech processing have been investigated for several decades, recent research has focused on investigating the possibility of decoding speech directly from neural activity. Communication impairments can originate from neuro-degenerative disorders that affect the motor production and articulation of speech, such as amyotrophic lateral sclerosis (ALS), as mentioned earlier, or from language disorders that affect the cognitive production or comprehension of language such as various forms of aphasia [12]. One goal of characterizing neural activity during speech production and comprehension is to develop neurotechnological applications to restore communication to those affected by speech and language disorders. The majority of neuroimaging studies of communication have used functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) to localize the neuroanatomy involved in speech and language processing [13], while noninvasive electromagnetic techniques such as magnetoencephalography (MEG) and electroencephalography (EEG) have additionally provided information related to the temporal patterns of activation. In particular, fMRI and PET have greatly contributed to the understanding of speech processing in the human brain. However, such techniques

that rely on measurements of hemodynamic responses (timescale of 4 - 6 seconds) are unable to capture the rapid temporal dynamics of natural speech (phoneme productions are often < 200 ms). ECoG, the measurement of electrical activity directly from the cortex, has become a highly promising neural signal acquisition modality for studying speech and language processing due to its capability to provide high spatial and temporal resolution [10]. Ultimately, the unique information offered by ECoG can be used to develop neuroprostheses that will decode intended speech for expressive communication or represent perceived language information directly using neural activity to augment receptive communication.

2.1 REVEALING THE NEURAL CORRELATES OF LANGUAGE

Prior to the development of functional neuroimaging techniques, identification of language areas in the human brain was based on studies of deficits in patients with damaged brains or patients undergoing electrical stimulation during neurosurgery [14]. The process of identifying parts of the brain involved in language processing began as early as 1861, when neurosurgeon Paul Broca studied the brains of nine patients with lesions and concluded that the expressive language centers present in most humans are located in the posterior frontal lobe of the left hemisphere, in an area now known as Broca's area [15]. A decade later, another renowned neurologist, Carl Wernicke, discovered that the posterior part of the left temporal lobe is involved in the comprehension of language. This region is now known as Wernicke's area [16]. Other researchers have also developed functional models of speech that describe the speech-related neural areas and their functional significance [17, 18]. These models

have identified the functional network consisting of the pre-motor cortex, primary motor cortex, Broca's area, primary auditory cortex, Wernicke's area, and superior temporal gyrus (STG) to be involved in the planning and production of speech and in the perception of speech (see Figure 3).

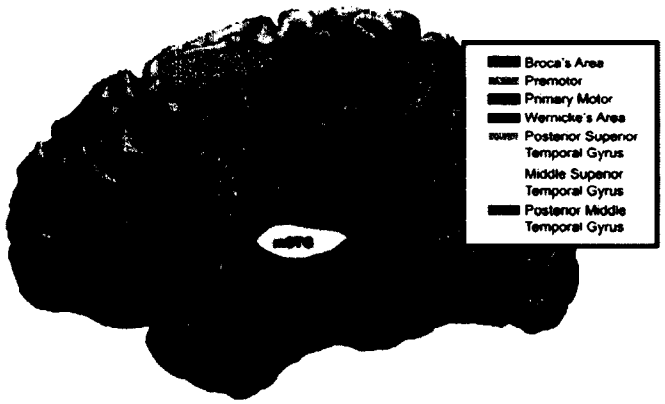


FIG. 3: The areas of the cortex involved in speech planning, articulation and production, shown on a generic brain using color maps demonstrating approximate locations of the different areas. The pre-motor area, primary motor area, and Broca's area are involved in speech preparation and articulation while the posterior and middle superior temporal gyrus, the posterior middle temporal gyrus and Wernicke's area are associated with speech perception and processing. Based on figure in [19].

Neuroimaging techniques such as PET or fMRI are also used to identify the neural correlates of speech processing by the human brain [13,20–24]. PET imaging involves the use of positron-emitting isotopes as tracers to detect changes in the cerebral blood flow and volume in response to a stimulus [25]. For speech processing, a stimulus might be the presentation of acoustic speech or the preparation and execution of a speech motor task. The spatial resolution in PET imaging is approximately 6 mm, while the temporal resolution is between tens of seconds to several minutes. In contrast, fMRI imaging results from changes in the blood oxygenation level due

to the metabolic activity of neuronal tissue without the use of a radioactive tracer. The spatial resolution of fMRI ranges from 1-4 mm, and the temporal resolution is between hundreds of milliseconds to seconds.

Studies have investigated the neural bases of speech production as well as the perception of speech using electroencephalography (EEG) [26–28]. Due to the comparatively low signal-to-noise ratio of EEG, time-locked averages known as event-related potentials (ERPs) are needed to capture the relevant brain responses. ERPs have been successfully used to study the temporal evolution of the phonological and lexical processes in the human brain corresponding to speech production and perception [26–28], but not during active speech production due to contamination from myoelectrical artifacts. While noninvasive measures such as EEG and MEG can theoretically offer adequate temporal resolution, factors such as spatial resolution on the order of centimeters, spectral bandwidth on the order of 80 Hz, and susceptibility to electrical artifacts severely limit the utility of these modalities for investigating the joint spatio-temporal dynamics of brain activity.

The excellent spatial resolution of fMRI and PET and the temporal resolution of EEG and MEG have provided theoretical and computational models that highlight the spatial topography and functional connectivity of the brain networks involved in speech production and comprehension [13, 26–29]. The results of these studies highlight the interconnectedness of neural networks for speaking, of which the traditional Broca's and Wernicke's areas play a crucial role in production and perception, respectively [13, 29]. Using intracranial electrophysiology, both the spatial and temporal

neurological dynamics of speech and language can be assessed simultaneously, providing a new opportunity to observe the entire speech network during production and perception tasks, which could not be done using standard noninvasive neuroimaging and electrophysiology. With the additional spatial and temporal detail available through ECoG recordings, it is now possible to develop a real-time neural prosthesis for speech and language.

2.2 THE ELECTROCORTICOGRAM

Signal Acquisition and Characteristics

ECoG measures the electrical activity of the brain recorded by electrodes placed directly on the surface of the cortex. ECoG was originally developed by neurosurgeons W. Penfield and H. Jasper in the 1930s as a technique for localization of epileptic seizure foci prior to surgical resection [30]. Modern ECoG typically uses platinum electrodes with a diameter of 4 mm that are implanted as either two-dimensional grids (e.g., 8x8 electrodes) or one-dimensional arrays (e.g., 4 or 6 electrodes) with an inter-electrode distance of 10 mm [31], as shown in Figure 4. In addition to standard clinical ECoG arrays, micro-ECoG arrays (center-to-center distance of 4mm or less) have also been used in recent studies to improve spatial resolution [32–35]. An example of a microgrid array is also shown in Figure 4.

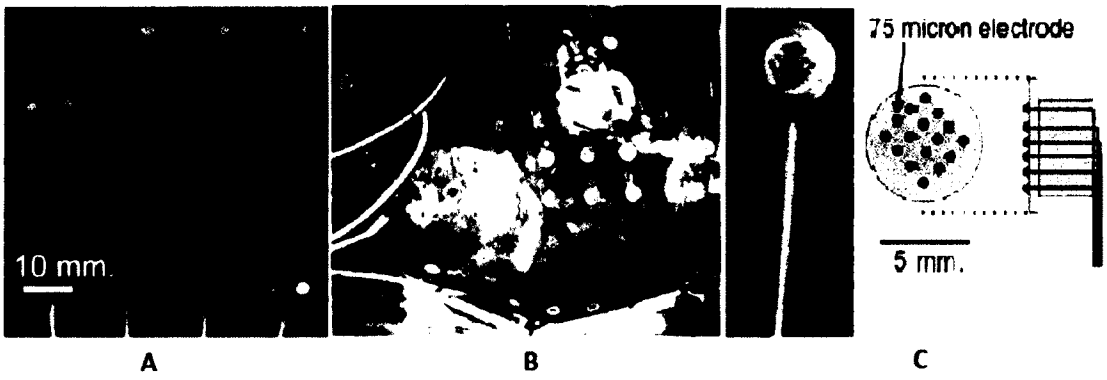


FIG. 4: Macro and micro ECoG arrays. A: Standard clinical macrogrid. B: Surgical placement of macrogrid. C: Microgrid array (left), schematic of microgrid array (right). Based on figure in [31].

ECoG recordings are well-suited for basic neuroscience research as well as for neural decoding studies. The recording characteristics of ECoG include: (1) spatial resolution on the scale of millimeters (activity is related to the neural tissue directly beneath the electrode disk), (2) frequency bandwidth up to 200 Hz or higher, (3) an amplitude up to 100 μV compared to near 20 μV for EEG, and (4) reduced sensitivity to movement and myoelectrical artifacts compared to EEG and MEG [31]. Due to higher signal amplitudes and lower sensitivity to artifacts, ECoG signals have higher signal-to-noise ratios (SNRs), which is highly desirable for any signal acquisition modality. Compared to penetrating electrodes, it is believed that ECoG does not suffer from adverse tissue reactions and electrode encapsulation issues [36] that can degrade signal quality over time because ECoG does not breach the cortex [37,38]. The superior bandwidth of ECoG is particularly important because brain activity in the 70-180 Hz gamma band range has been linked to perception, cognitive function, and motor tasks [4,39–42]; including learning, memory, and speech [19,43–64]. These

studies demonstrate the promising real-time decoding potential of ECoG compared to other neuroimaging modalities.

Signal Processing and Feature Extraction

To date, the most relevant information or features of ECoG are based on its spectral dynamics. ECoG recordings typically require preprocessing to condition the signals for further analysis. A spatial common average reference (CAR) filter is commonly applied to remove any low-frequency fluctuations and artifacts that may be present on all channels over a region [43,44,50,53]; then signals are high-pass filtered starting between 0.5-2 Hz to further reduce low-frequency fluctuations and heartbeat artifacts [51,53]. The signals are also notch or comb filtered at harmonics of 60 Hz (or 50 Hz as appropriate) to eliminate power line interference [47,49–53]. Furthermore, any trials or channels that show excessive fluctuations, presence of outliers, low SNR, or are overlying pathological tissue are generally removed to ensure that these trials or channels do not bias the analysis [47,51,53].

Following the pre-processing stage, more advanced signal processing techniques are used to extract the relevant spectro-temporal features for further analysis. The signals are typically transformed to the frequency domain using a Discrete Fourier Transform (DFT), auto-regressive model, or band-power filtering [43,44,50]. After this, the power of the signal is determined over the frequency bands of interest by some form of averaging of spectral amplitudes in the respective frequency ranges. To date, the gamma band, from approximately 30 Hz to 200 Hz, has provided the most informative description of the neural processes underlying speech. Within this

frequency range, modulations of the high-gamma band, from approximately 60-80 Hz to 150-200 Hz, have been identified as highly correlated with speech production and perception. This range of frequencies is of particular interest in ECoG because it is not detectable with the limited spectral ranges of other neural signal acquisition techniques such as scalp EEG [31]. Nevertheless, ECoG is also well suited to examine the delta (<4 Hz), theta (4-8 Hz), alpha (8-12 Hz) and beta bands (12-30 Hz) in the context of speech production and perception in the human cortex [19, 43] and often results in superior signal quality due to the reduction of electrical artifacts compared to EEG.

The extracted spectro-temporal ECoG features have been used to both study and decode neural activity during speech production and perception. For spatial characteristics studies investigating the speech network, recorded signals from electrodes at different locations over the cortex are compared against a reference signal (e.g., recorded speech) using statistical techniques such as correlation analysis and/or analysis of variance (ANOVA) [19,43,44,47,50,52–55]. To obtain the temporal dynamics of these spatial networks, ECoG signals from each channel are compared to the reference signal at different time latencies relative to the onset of speech articulation or the presentation of speech stimuli to quantify the neural processing involved in the production and perception of speech, respectively [50, 52]. Decoding articulated or perceived speech from ECoG generally requires the application of additional advanced signal processing and machine learning techniques to produce the desired speech outcome. Some recent attempts at speech decoding using these approaches

are discussed in Section 2.4.

2.3 NEURAL DYNAMICS OF SPEECH AND LANGUAGE PROCESSING USING THE ELECTROCORTICOGRAM

Numerous studies have investigated cortical activity using ECoG during various speech tasks to identify the cortical areas and networks involved in speech production and perception, which create the groundwork for speech decoding. The following sections summarize the studies focusing on the identification of the spatial and temporal dynamics of ECoG during speech production and perception.

Spatial Characterization of Speech Production

A recent study by Bouchard et al. (2013) examined modulations of the ECoG high gamma-band during production of consonant-vowel syllables to investigate the phonetic organization of the speech sensorimotor cortex [48]. Spatial patterns of cortical activity showed that the gamma band activity, recorded by electrodes over the sensorimotor cortex, demonstrated different spatial organizations for consonants versus vowels. The spatial patterns also confirmed prior neuroimaging findings that speech is produced through the coordination of a distinct set of articulatory representations in the ventral sensoriomotor cortex. In another study, Pei et al. (2011) analyzed the high-gamma power of ECoG signals recorded while subjects overtly or covertly repeated words presented acoustically or visually [44]. Overt word production was associated with high-gamma power changes in the superior and middle parts of the temporal lobe, Wernicke's area, Broca's area, the pre-motor cortex and

the primary motor cortex. Covert word production, in contrast, was associated with high-gamma changes in the superior temporal lobe and the supramarginal gyrus. This study provided corroborating evidence for an overt speech production network identified in prior neuroimaging studies [13], but conflicted with prior accounts of covert speech, which merely suggested that the speech network was reduced in size and strength for the covert condition as compared to the overt condition [13, 21]. This study also demonstrated weaker and less distributed cortical activations during the covert speech condition, but in certain areas, in particular, the superior temporal lobe, no significant difference in activation was found between the overt and covert conditions. This study, thus, highlights the important role played by the superior temporal lobe during covert speech production. The neural correlates of verb generation and noun reading have also been investigated using an analysis of the ECoG high-gamma band [49]. The results of this study showed that activation was found in the primary mouth motor area, the superior temporal gyrus (STG), and Broca's and Wernicke's areas, which agrees with previously identified regions involved in speech production.

Spatial Characterization of Speech Perception

ECoG has also been used to explore the differences between the processing of speech and non-speech auditory stimuli, such as tones, in order to primarily highlight the importance of the human cortex in processing both the acoustic and the phonological aspects of speech. One of the early ECoG speech studies based on

speech perception by Crone et al. (2001) explored the temporal and spatial activations of the cortex in response to perception of speech (phonemes) and non-speech (tones) stimuli [43]. This study found that activations in the primary auditory cortex and STG occur in the gamma band, and to a higher extent in the high gamma band, during phoneme discrimination. For the tone stimuli, it was found that increases of the gamma power occurred in fewer electrodes, i.e., to a smaller spatial extent, and with lower magnitudes than for phoneme stimuli. This effect was particularly noticeable in the left STG, which highlights the importance of the left hemisphere's auditory cortex in speech processing as compared to non-speech auditory processing, and confirms results from earlier lesion studies that investigated tone perception.

Additional studies have supported the claim that processing of incoming auditory stimuli results in an increase in high-gamma activations in the left STG [51], and have elucidated the importance of the speech envelope in speech comprehension [50, 52]. The speech envelope is the rectified speech waveform that fluctuates with speech intensity, i.e., loudness, phonetic content, and rhythmic cadence that is vital for understanding fluent, conversational speech. ECoG has the requisite spatial and temporal resolution to study this quickly changing speech signal, and prior work has found that ECoG in the belt areas of the auditory cortex, i.e., the areas lying relatively early in the auditory pathway, tracks modulations in the speech envelope as well [50]. This provides evidence that cortical signals closely represent the acoustic features of speech, and paves the way for future studies to investigate finer temporal aspects of speech processing in the cortex. An important study by Canolty et al.

(2007) showed that the high-gamma activity in the ECoG tracked the spatio-temporal dynamics of word processing while subjects listened to a stream of verbs associated with actions of the hand and the mouth [53]. From this study, it was found that the perception of verbs activates the posterior and middle STG as well as the superior temporal sulcus (STS), which supports previous studies that found evidence that the STG is largely involved in speech comprehension.

Other studies have attempted to investigate the cortical responses to altered speech feedback to identify the neural dynamics of sensory processing for error detection and correction. One particular study by Chang et al. (2013) recorded speech from subjects, and then later played back these recordings with slight perturbations in the pitch to the subjects while they were speaking [54]. It was found that the cortical responses in the posterior STG were suppressed while listening to unaltered feedback, but enhanced in response to the pitch-altered feedback, which corroborates results from a previous study which demonstrated the same effect in the EEG auditory response [65]. With ECoG it was possible to localize this change directly to the auditory cortex, while EEG only provides indirect evidence of auditory cortex involvement. The subjects were found to compensate for the altered pitch in the stimuli by changing the pitch of their speech. Furthermore, these vocal changes made by the subjects were predicted by their auditory cortical responses to the altered pitch stimuli. This neurological relationship provides evidence for the sensorimotor control of articulation in humans through the coordination of various cortical areas.

Temporal Evolution of Speech: From Planning and Production to Perception

Recent studies have attempted to use language tasks that involve both speech production and perception to simultaneously analyze both expressive and receptive speech areas [55–57]. It can be concluded from the combination of all these studies that, broadly, different areas of the cortex are activated by motor speech production and speech perception, as illustrated in Figure 3 (Page 10). The areas involved in speech articulation are the pre-motor cortex, which is mainly involved in planning; the face-mouth-motor regions, involved in generating mouth movements necessary for speech articulation; and Broca's area which is involved in speech planning and articulation. The areas of the cortex primarily involved in speech perception and comprehension are STG and Wernicke's area. Analysis of the temporal dynamics of ECoG signals during speech perception shows that the posterior STG is activated first, followed by the middle STG, then STS [53]. The spatial characterization results found by these studies analyzing both speech production and perception simultaneously, further support results from the studies investigating speech production and speech perception independently, discussed in the previous two sections, respectively.

2.4 ECOG-BASED DECODING

The aforementioned studies combined provide a framework for associating ECoG recordings (namely the high-gamma band power) with the behavioral tasks of speech production and perception. Figure 5 illustrates the concept of training a neural-based

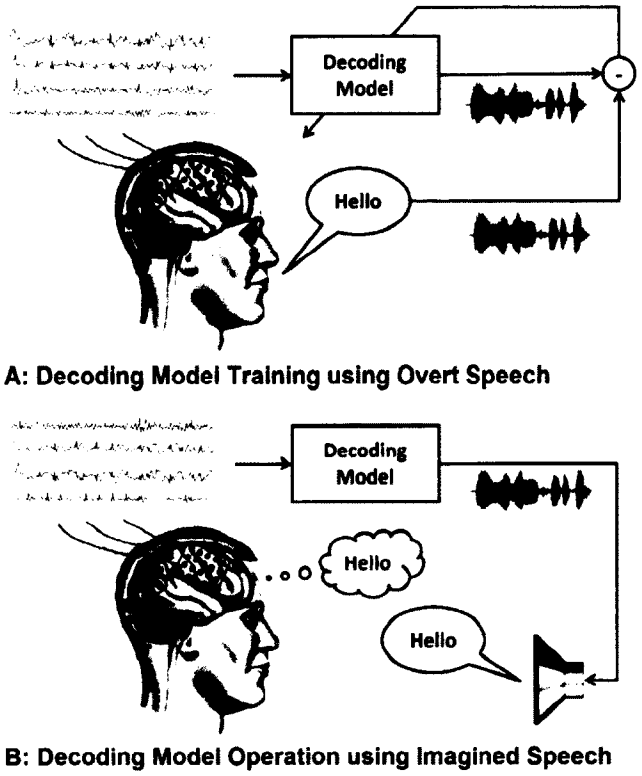


FIG. 5: A typical speech decoding model. (A) ECoG signals serve as the input to a neural-based decoding model that is trained using representations of recorded overt speech (e.g., time series, spectrogram, etc.) to ideally reconstruct the overt speech directly from the ECoG signals. (B) In principle, a version of the trained decoding model would be used to generate imagined speech directly from ECoG signals in real time. The model may be trained using overt speech or through some prior characterization of imagined speech, since there is no behavioral output during imagined speech.

decoding model using overt speech for reconstruction of imagined speech. This concept represents the basis of a speech neuroprosthetic device. The decoding model may perform a continuous reconstruction of the speech or a discrete classification and output of phonemes, words, etc., depending on the objective and constraints of the system. While it may be possible to decode individual words or phrases discretely, extending such models becomes highly dependent on the desired vocabulary and can

lead to a combinatorial dilemma. Alternatively, the ability to decode formants or phonemes will enable the creation of generative models that are not limited to a fixed vocabulary. In terms of decoding perceptual information, one potential application lies in the development of a practical auditory neuroprosthesis. A practical auditory decoder should demonstrate the ability to segregate and process attended sound streams from irrelevant signals from more than one point in space in a complex acoustic environment, such as in a hospital setting or a restaurant. This ability would expand the active space of the signal, allowing for more natural communication when competing multi-talker streams exist in the acoustic space. Additionally, perceptual decoding could be used to enhance the speech production decoder by accounting for the production-perception feedback loop [54,56]. The following sections discuss attempts to decode, or predict speech behavior directly from ECoG activity. These studies represent the initial steps toward the development of real-time neuroprostheses for speech and language.

Decoding of Overt Speech Production

Speech production is a complex process that is initiated by linguistic processing and results in articulated speech, which can be further broken down into sentences, words, syllables, vowels, consonants, and phonemes. The following speech decoding studies have investigated the decoding of these various levels of speech production. A study by Wang et al. (2011) showed an increase in gamma power over Broca's and Wernicke's areas during picture naming and property identification [57] that was used to decode the semantic category associated with each stimulus. In this study,

ECoG was recorded from four subjects as they participated in a picture naming task, where they were presented with pictures of objects from different semantic categories, such as food, tools, dwelling, and body parts, varying from subject to subject. Two popularly used machine learning techniques, the Gaussian Naive Bayes Classifier and the linear support vector machine (SVM), were then used to decode the semantic category that the subjects named from the ECoG activity, resulting in accuracies as high as 74% (chance level 33%).

A study by Kellis et al. (2010), which involved the use of micro-ECoG electrodes implanted over the facial motor cortex and Wernicke's area, attempted to classify 45 possible pairs among a set of 10 words that a subject was articulating. The power spectra of the ECoG data and principal component analysis (PCA) were used to maximize the variance between the classes of words being identified [32]. The electrodes that led to the best classifier performance were selected to improve classification of word pairs, which demonstrated that electrodes implanted over the face-motor cortex resulted in better classification (40 of the 45 word pairs classified with an accuracy of 80% or higher, chance level 50%) than did the electrodes implanted over Wernicke's area (15 of the 45 word pairs classified with an accuracy of 80% or higher). This may be explained by the role of the face-motor cortex in the control of mouth movements required for speech articulation.

A study by Pei et al. (2011) examined ECoG signals to classify four vowels and nine pairs of consonants from a closed set of spoken whole words, and achieved 10-fold cross-validation classification accuracies up to 43% for vowel and 49% for consonant

pairings, which were both significantly better than chance (chance level 25% for both) [58]. This study used the technique of maximum relevance and minimum redundancy (MRMR) to select the top 35 or 40 ECoG features for decoding consonant pairs or vowels, respectively, using the Naive Bayes Classifier. A recent study by Kanas et al. (2014) performed spatio-spectral feature clustering of ECoG recordings in order to detect speech activity from one subject during a syllable repetition task, achieving an accuracy of 98.8% [59]. This study used the method of k-means clustering to group the best power spectral density features for all the channels and frequencies into clusters. This grouping was followed by the classification of the ECoG features using different algorithms, among which the support vector machine was the most accurate in detecting speech activity from the related cortical signals. A study by Zhang et al. (2012) used a sentence-level approach to classify ECoG high gamma responses obtained from the posterior portion of the inferior frontal gyrus during the production of two eight-character Chinese sentences with a 77.5% accuracy of classification (chance level 50%) [60]. Dynamic time warping was used to align the ECoG responses with the onset of sentence articulation and identify temporal activation patterns, or ECoG response templates, for the two sentences. Fisher's linear discriminant analysis (LDA) was used for classification of the two sentences based on the time-warped ECoG responses, which was then evaluated using a leave-one-out cross validation technique. This study demonstrates the discriminability of high gamma activity recorded during speech at the sentence level, which is an important

advance toward a neuroprosthesis for fluent, conversational speech. The studies discussed up to this point have attempted to decode semantic categories, words, vowels, consonants, and sentences in articulated speech from ECoG activity, and have shown preliminary success in speech decoding. However, the success rates for these decoding techniques are still low for a speech prosthesis capable of operating in natural environments where both decoding speed and accuracy are of importance.

One possible way to improve the information rate of speech decoding may be to predict the smallest identifiable components of speech, called phonemes, which can be sequenced together to form more complex productions. A study by Blakely et al. (2008) used micro-ECoG grids to successfully classify a set of four phonemes, in a pair-wise fashion, using ECoG high-gamma power [33]. The ten best channels for classification of each of the phoneme pairs were found using a correlation-based feature selection technique, followed by a binary classification with a linear support vector machine, which was then validated using a 4-fold nested cross-validation. The study found that different locations on the cortex were specific to classification of particular phoneme pairs, thus demonstrating the spatial separation of phoneme representation in the human brain. Using the best set of electrodes for each phoneme pair, accuracies as high as 75% for classification of the "RA" versus "LA" pair and 70% for the "BA" versus "WA" pair (chance level 50% for both), were achieved. Extending these possibilities further, a recent study by Mugler et al. (2014) investigated a technique to decode the entire set of phonemes in American English using ECoG recordings from four subjects while they produced words from the modified rhyme

test. This test consists of 300 words with similar frequencies of phoneme occurrence as found in the English language [61]. ECoG feature selection was performed on the time-frequency features (short time Fourier Transform features in the mu, beta and high gamma frequency bands) using an ANOVA and selecting features as those with the lowest p-values. These features were then used to classify phonemes using the linear discriminant analysis technique followed by a ten-fold cross-validation to evaluate the classification performance. For the subject with the best performance, 36.1% of all consonant phonemes (chance level 7.4%) and 23.9% of all vowel phonemes (chance level 12.9%) were correctly classified, with a classification rate as high as 63% for a single phoneme. Another study by Leuthardt et al. (2011) was able to successfully classify the production of two phonemes based on the squared correlation of the ECoG high gamma power during overt [35]. This was an online study where two subjects produced two different phonemes to control a one-dimensional cursor on the computer screen. The cursor was controlled using a weighted summed value, based on the decoded phoneme (e.g., one phoneme moved the cursor right, the other left). Classification rates of 76% and 91% were achieved for the two subjects (chance level 46.2%).

Decoding of Imagined Speech Production

Because the primary goal of a speech neuroprosthesis is to restore communication to those who are able to achieve little or no normal verbal communication, it is vital to demonstrate that imagined or attempted speech can be accurately decoded from brain activity. Pei et al. (2011) examined ECoG recordings to classify vowels and

consonant pairings during covert word repetition, achieving classification accuracies as high as 43% for imagined vowel classification and 46% for consonant pairs (chance level 25% for both) in imagined speech [58]. This was done using the same technique for feature selection (MRMR), classification (Naive Bayes Classifier), and evaluation (10-fold cross-validation) as used for the decoding of overt vowels and consonant pairs. This was one of the first studies to demonstrate the possibility of classifying different vowels and consonants embedded in imagined words directly from brain signals. Furthermore, the decoding results were similar for both actual and imagined speech, which provides evidence that imagined speech can also be decoded from neural activity. Leuthardt et al. (2011) also investigated online control of a one-dimensional cursor using ECoG high-gamma power for an imagined phoneme versus rest task in two subjects (e.g., imagination of a phoneme moved the cursor right, rest moved it left) [35]. The feature selection and classification techniques used were consistent for the overt and covert phoneme pair identification (discussed in the previous section). This resulted in closed-loop classification accuracies above 69% (chance level 42.6%) for both the overt and covert conditions, with the overt condition yielding a performance as high as 91%.

A recent study by Martin et al. (2014) explored the possibility of predicting spectro-temporal components of imagined speech from ECoG high gamma power, in a manner similar to overt speech [62]. In this study, subjects read aloud (overt condition) and imagined reading (covert condition) short stories that scrolled across the computer screen. Neural decoding models were then developed for the overt speech

condition in order to predict two speech feature representations: (1) a spectrogram-based feature, which is a time-varying speech amplitude envelope at different acoustic frequencies, and (2) a modulation-based feature, which is a non-linear transformation of the spectrogram. Linear decoding models were developed in order to predict the two speech representations from ECoG high-gamma activity for the overt condition. These models were then applied to predict the speech representations during the covert condition. Dynamic time warping was used to align the reconstructed covert speech representations to the actual overt representations. The correlation coefficients between the actual and predicted speech representations for the overt condition, and between the time-warped actual and predicted speech representations for the covert condition, were used to evaluate the models. The reconstruction correlation was found to be statistically significant in all the subjects for the overt condition. For the covert condition, the predictions were statistically significant when compared to the baseline condition. This indicates that auditory representations of imagined speech can be reconstructed from models developed for actual speech, showing that both overt and covert conditions share a common neural basis.

Decoding of Speech Perception

Other studies have investigated the possibility of decoding perceived speech directly from cortical recordings. A study by Zavaglia et al. (2012) analyzed auditory features to build a forward model of the ECoG responses corresponding to word and acoustically matched non-word stimuli presented to the subject [63]. This is done by using a weakly-coupled oscillator model of transient synchronization (WCO-TS). The

WCO-TS uses the auditory stimulus being presented to the subject as the input and utilizes the serial nature of word processing in the human cortex, as demonstrated in [53], in order to predict the ECoG gamma activity corresponding to incoming auditory stimuli. Although this is inverse to the process of speech decoding, i.e., utilizing speech features to predict neural information, it provides useful information that may be analyzed to build a direct model for the prediction of speech features from ECoG. This study also identified a set of speech features called the "occurrence time" features which were found to outperform standard cepstral features typically used in speech recognition, especially in noisy recognition environments. These occurrence time features correspond to the occurrence of peaks in specific speech frequency bands and may be useful in future decoding efforts.

Chang et al. (2010) measured ECoG activity in the posterior STG using a high density micro-ECoG grid during the presentation of three consonant-vowel syllables [34]. The study found that an acoustically varying speech stimulus is transformed into distinct phoneme categories in the human cortex. Using this information, it was possible to classify three consonant-vowel syllables from the ECoG signals recorded across the posterior STG. The dissimilarities between the neuronal response patterns were determined using a multivariate pattern classifier which uses L-1 norm regularized logistic regression, whose classification measures were used to construct a confusion matrix for each time interval. Multi-dimensional scaling of this confusion matrix and k-means were then used to classify the neuronal responses into three categories that corresponded to the three phonemes. The results from this study indicate

that the posterior STG performs a critical role in the phonological processing and categorization of perceived speech.

A significant contribution for perceived speech prediction was made by Pasley et al. (2012), which examined ECoG recordings from the superior temporal gyrus to reconstruct the speech spectrogram of aurally presented words and sentences [64]. Two representations for the perceived speech were found, similar to those used in [62], i.e., a spectrogram-based representation and a non-linear modulation-based representation. Linear neural decoding models were then developed which used the ECoG high-gamma power to predict these two speech representations, leading to linear and non-linear models respectively. It was found that slow and moderate time modulations in the speech, such as syllable rate, were reconstructed well using the linear model, i.e., these modulations are well-represented with the spectrogram-based representation. Fast temporal modulations, such as syllable onsets and offsets, could be better predicted using the non-linear model, i.e., they are well-represented with the modulation-based representation. The fidelity of the reconstructions of the spectrogram were sufficiently accurate to identify individual words directly from the reconstructed spectrogram-based speech representations using a simple spectrogram matching algorithm, leading to a median word identification percentile rank of 0.89 for 47 words (chance level 0.50).

2.5 LIMITATIONS OF EXISTING ECOG-BASED SPEECH STUDIES

These collections of studies have identified important neural correlates associated

with speech production and perception. Decoding models have also been successfully developed that are capable of predicting the essential components of verbal communication, namely the production and perception of speech and language directly from cortical activity. ECoG studies have been especially informative in their ability to provide a spatio-temporal characterization of the neural correlates of speech planning, production, and perception in the human cortex. These studies have specifically focused on the ECoG gamma-band power recorded over language-related cortical areas, which is significantly correlated with speech processing and is useful for speech decoding algorithms. However, the performance of ECoG-based speech decoders are not nearly as robust as needed for practical ECoG-based speech reconstruction, which is the ultimate aim of this type of research. One focus of future ECoG-based speech studies should include improving computational models of neurological speech processing for more accurate decoding. New models will benefit from continued research on the development of more advanced signal processing and ECoG electrode design to capture the recorded signals with higher fidelity. Most of the studies described in this review have implemented relatively simplistic linear approaches to characterize and predict speech components from cortical activity. However, in reality, it is likely that the relationship between ECoG activity and the various speech representations of interest are highly non-linear and dynamic. Therefore, more sophisticated models based on improved signal acquisition capable of capturing such non-linear relationships need to be developed to achieve a more transparent and practical ECoG based speech decoder.

The majority of speech and language ECoG studies have been performed using relatively discontinuous speech production and listening tasks, such as cued word repetition tasks. These studies are critical for identifying baseline neurological activation during speech, but are not fully representative of fluent, conversational speech. A characterization of the neural correlates associated with continuous and spontaneous speech production and perception may provide the supplementary information needed to develop more advanced models. The ability to decode perceived speech and articulatory commands in continuous and fluent communication will represent a fundamental improvement in the potential impact of a neural prosthesis for speech. Natural verbal communication often takes place in the presence of background speech or environmental noise. The perception and comprehension of speech in background noise requires additional verbal working memory and attentional resources to process the target stream and segregate it from competing background noise [66–68]. In addition, the auditory-motor feedback loop which is active during vocalization, has been shown to affect speech production [54, 56]. However, the potential interference of background noise on the feedback loop during communication is not well understood, particularly at low signal-to-noise ratios, and certainly not included in current decoding algorithms. Future ECoG speech studies will need to determine the contribution and coordination of both specific speech and non-speech regions to the production and processing of speech in different levels of background noise.

Furthermore, most of the ECoG-based speech decoding studies have focused on

decoding speech features such as the envelope, words, phonemes, vowels, and consonants from ECoG activity. However, modern real-time speech processors and automatic speech recognition (ASR) systems employ many other speech features, such as the formant frequencies, linear predictive coding coefficients, and mel frequency cepstral coefficients, among others [69]. While some studies have used formants for ECoG-based speech decoding [34], most other speech representations used in ASR have not been investigated with regards to their relationship to cortical activity. The decoding of these speech representations directly from ECoG activity is a practical next step, and the results could then be used in speech synthesis and recognition systems. This research would develop a natural extension of ASR to neurological data, and provide a step toward neural speech prostheses.

2.6 SPEECH REPRESENTATIONS USED IN SPEECH ANALYSIS AND PROCESSING

Speech processing is an important area of research, where certain properties of speech signals are extracted through various processing pipelines and used for different applications. This usually requires transforming the speech signal into a set of parameters, or a collection of signals, for the purpose of data reduction and parameterization [69]. These reduced representations contain the relevant information in the speech signal in an efficient manner. Different representations are typically useful in different applications, such that the useful representation retains the relevant information for a particular application, while removing information which is irrelevant to it.

Certain speech representations are based on understanding how speech is produced in the human vocal system. Speech can be thought of as an electrical signal, which consists of an envelope, a periodicity, and a fine structure [70]. The envelope of speech has been found to be important for speech comprehension as the manipulation of the speech envelope has been found to affect the recognition of vowels, consonants and sentence comprehension[71]. It has also been observed that human listeners can understand speech with a preserved temporal envelope but degraded spectral information [72]. The periodicity in a speech signal is reflected in the fundamental frequency, which is a time-varying frequency, evaluated over short windows of time and is qualitatively equivalent to the pitch. This fundamental frequency is an important speech feature which is vital for speech recognition and speaker recognition [70, 73]. Other than the fundamental frequency, other harmonics exist in the speech signal, reflected as peaks in the speech spectrum which also coincide with the resonances in the vocal tract [74]. Formants are popularly used in speech recognition because the information that humans require to distinguish between vowels is supposed to be contained purely in the frequency content of vowels, often characterized by the formants [75]. A slightly different line of research models the speech signal as a combination of a sound source and a vocal tract filter, commonly known as the source-filter model [69]. The sound source is assumed to produce an impulse train, with period equal to the pitch period or frequency equal to the fundamental frequency, for voiced sounds, and white noise for unvoiced sounds. The vocal-tract filter is approximated as an all-pole filter, whose coefficients are determined by linear

prediction or autoregressive modeling, and are known as the linear prediction coefficients (LPC) [69]. The LPC coefficients, in combination with the pitch, can be used to encode the important information in human speech, and are important for speech recognition and reconstruction [76].

Some other speech representations are based on the understanding or knowledge of how speech is perceived by the human auditory system. One of the most important speech representations in this direction is the mel frequency cepstral coefficients (MFCC). MFCC is a speech representation based on mapping the speech spectrum onto a non-linear mel scale of frequency by passing it through a bank of non-linearly spaced filters. The mel scale, or the melody scale, is a logarithmic scale based on how the human ear perceives pitch. MFCCs are commonly used for speech recognition, speaker recognition and speech reconstruction [77, 78]. Another important speech representation, known as the Perceptual Linear Prediction Coefficients (PLP), is often used in speech recognition, either alone or in combination with other features [79,80]. The PLP coefficients are obtained by mapping the power spectrum of speech onto another non-linear scale, known as the Bark scale, which is a scale based on the perception of loudness by the human ear. Some other speech representations use filter banks based on the human cochlear processing of speech, in order to filter speech prior to feeding it into a speech recognition system [81–83]. These cochlear filter banks are based on functional models of the cochlea and are supposed to preserve important information in both time and frequency domains. They are also supposed

to be more robust for speech processing than some other representations [82]. However, the processing of speech based on the human cochlear filter-bank also requires more complex computation than some other techniques; thus, there is a trade-off between robustness and computational efficiency when using these features.

CHAPTER 3

EXPERIMENTAL METHODOLOGY

This chapter describes the methodology used for the experimental paradigm and data acquisition. The signal pre-processing as well as the primary techniques used for characterization and development of the decoding models are described.

3.1 DATA ACQUISITION

Data were collected from eight patients with medically intractable epilepsy who were undergoing treatment at the Albany Medical Center. The subjects underwent temporary placement of subdural electrode arrays to localize seizure foci prior to surgical resection of the epileptic tissue. All the eight subjects gave informed consent to participate in this study, which was approved by the Institutional Review Board of the hospital. All of the subjects were mentally, visually and physically capable of performing the task and had performance IQs of 85 or higher.

The implanted electrode grids, produced by Ad-Tech Medical Instrument Corporation (Racine, Wisconsin), consisted of platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) that were embedded in silicone and had an inter-electrode distance of 1 cm. One subject (Subject H) had an electrode grid with an inter-electrode distance of 6 mm. Grid placement and duration of ECoG monitoring was based purely on the clinical requirements of the subjects without any consideration for this study.

Each subject had postoperative anterior-posterior and lateral radiographs, as well as computed tomography (CT) scans to verify grid locations. Three-dimensional cortical models of individual subjects were generated using pre-operative structural magnetic resonance imaging (MRI). These MRI images were co-registered with the post-operative CT images using Curry software (Compumedics, Charlotte, NC) to identify electrode locations, as shown in Figure 6. Cortical locations were derived using Talairach's Co-planar Stereotactic Atlas of the Human Brain [84] and a Talairach transformation (http://www.talairach.org). Cortical activation maps were generated using custom Matlab software. Activation maps computed across subjects were projected on a three-dimensional cortical template provided by the Montreal Neurological Institute (MNI) (http://www.bic.mni.mcgill.ca).

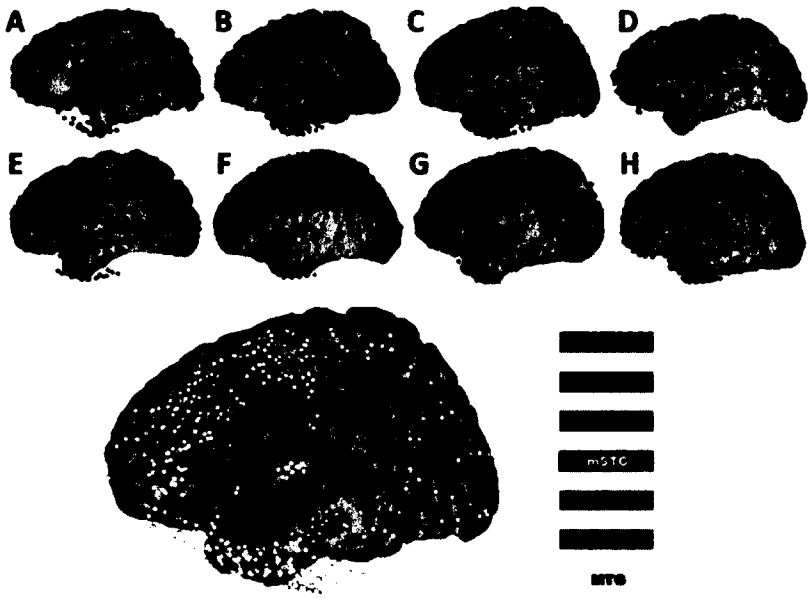


FIG. 6: (Top Row) Placement of electrodes on all 8 subjects. (Bottom Row) Combined electrode placement over all 8 subjects. Electrode locations were identified in a post-operative CT, co-registered to pre-operative MRI, and transformed into a joint Talairach space for comparison across subjects.

3.2 TASK AND DATA COLLECTION

In this study, each subject was seated in a semi-recumbent position in a hospital bed, about 1 m away from a computer screen. During the experiment, the text of a famous passage, either a political speech (The Gettysburg Address given by President Abraham Lincoln during the Civil War in 1863 or President John F. Kennedy's inaugural address in 1961), a nursery rhyme (Humpty Dumpty), or a fictional story (Traitor Among Us: Charmed, an adventure/supernatural story written for the TV show Charmed), ranging from 109 to 411 words, scrolled from right to left across the screen at a constant rate between 20% and 35% per second, resulting in run durations between 129.87 and 590.10 seconds. For each subject, this rate was chosen to be a rate that the subject was comfortable with, based on his/her attentiveness and cognitive/verbal ability. Only one of three passages were visually presented, to each subject, during the experiment. For all the subjects, ECoG was recorded in two different conditions, one in which the subject was instructed to read the presented text out loud, and the other in which the subject had to read the presented text silently, i.e., overt and covert tasks, respectively. Each overt run was followed by a covert run with the same text scrolling across the screen at the same rate as for the overt task. In this dissertation, results from the overt task are presented.

The experimental paradigm is shown in Figure 7. ECoG signals were recorded at the bedside of the patients using eight 16-channel g.USBamp biosignal acquisition devices, designed by g.tec (Graz, Austria). ECoG signals and the speech signal recorded from the microphone, while the subjects were speaking, were simultaneously

digitized at a sampling rate of 9600 Hz. Electrode contacts which were distant from the epileptic foci and areas of interest were used as reference and ground electrodes. The recordings were visually inspected offline for environmental artifacts and interictal activity. Channels which did not contain any ECoG activity were removed before the analysis, which resulted in 56-120 channels each for the eight subjects. Along with recording cortical activity, the subjects' eye gaze was recorded using a monitor with a built-in eye tracker, designed by Tobii Technologies (Stockholm, Sweden). The eye tracker was calibrated in a subject-specific manner prior to the experiment using custom software. Data collection from the signal acquisition devices, microphone, eye tracker and the control of the experimental paradigm was done simultaneously using the BCI2000 software [85].

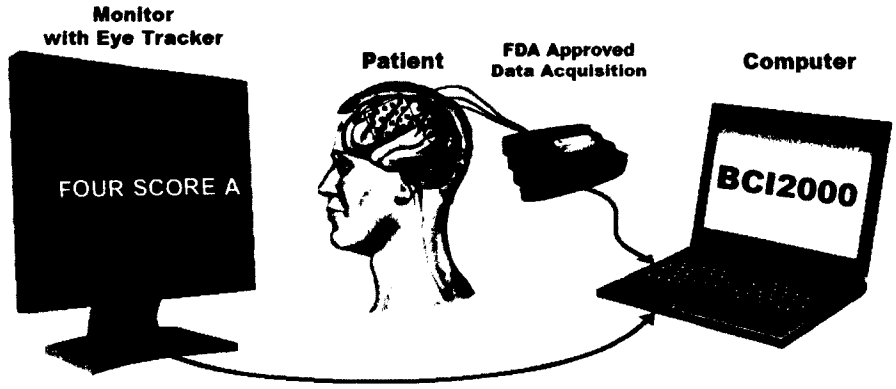


FIG. 7: The experimental paradigm: The subject was presented with scrolling text on a computer screen with a built-in eye tracker, which verified the location of eye gaze on the screen during data acquisition. The eye tracker and data acquisition devices were interfaced with a computer running BCI2000.

3.3 DATA ANALYSIS

Pre-processing

The raw ECoG signal from each electrode was first high-pass filtered with a cutoff frequency of 0.01 Hz, to remove low-frequency or dc components. After removing channels without any ECoG activity, the remaining channels were spatially re-referenced using a common average reference (CAR) montage, i.e., by subtracting from each channel the average of the ECoG signals over all the channels. The resulting signals were low-pass filtered and decimated to 400 Hz. A finite-impulse-response (FIR) notch filter in the range 116-124 Hz with zero-phase (forward and inverse) and -60 dB stop-band attenuation was used to remove the 120 Hz power line interference harmonic. This was done because the frequency range between 70 Hz and 170 Hz was used as the gamma band, and 120 Hz falls within this range. The 60 Hz power line interference was not notched out because none of the frequency bands considered for ECoG activity included 60 Hz. Following this, three band-pass filters with zero-phase and -60 dB stop-band attenuations were computed as follows:

(a) Mu-band filter: 8 Hz-12 Hz

(b) Beta-band filter: 13 Hz-26 Hz

(c) High Gamma-band filter: 70 Hz-170 Hz

These band-pass filters were applied to the notch-filtered ECoG signals to extract ECoG activity in the mu band, the beta band, and the high gamma band respectively. Following this, the envelopes for the mu, beta, and gamma activities and the speech

activity were computed as the squares of the magnitudes of the analytic signals obtained from the Hilbert Transform [86]. The Hilbert Transform is a linear operator used commonly in signal processing to derive the analytic representation of a signal. It is related to the Fourier Transform by the following equation:

$$F(H(u))(\omega) = (-i\ sgn(\omega))F(u)(\omega) \tag{1}$$

The Fourier Transform could also be used for power or envelope computation. However, the ECoG signal is highly non-stationary, and since we are computing the instantaneous power envelope, the Hilbert transform gives a high temporal resolution and captures the rapid changes in the amplitude of the signal. Furthermore, since the ECoG signal can be considered narrow-band for small time durations, the Hilbert Transform is useful for computing its power envelope. The data was then low-pass filtered using an eighth-order low-pass Chebyshev Type I filter with a cut-off frequency of 8 Hz, and decimated to 20 Hz.

Characterization

The primary measure used here for characterizing the relationship between the ECoG power at a particular location and the speech power was the Pearson's correlation coefficient at six different latencies, between -300 ms to +300 ms, spaced by 100 ms, and the p-value or the statistical significance of the resulting correlation coefficients. The Pearson correlation coefficient between two signals, X and Y, of sample sizes n, is given by:

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} \tag{2}$$

The statistical significance of the relationship between the neural signals and the auditory signals was assessed by calculating the correlation coefficient between these two signals in each cortical region, and finding out how far away this distribution of correlation coefficients is from zero. This was done using a bootstrapping randomization test in which the ECoG power samples for each channel were flipped and then circularly shifted by a random amount to preserve their statistical properties while destroying their temporal properties. For each channel, this was repeated 1000 times, and the Pearson correlation coefficients were used to form a beta distribution. Then, the actual Pearson coefficient between the actual or unscrambled ECoG signal and the auditory signal was fit to this distribution to find the p-value. In the spatio-temporal plots, the negative logarithm (base e) of this p-value was plotted across all the channels, for time lags ranging from -300 ms to 300 ms, in steps of 100 ms. The p-values were then Bonferroni-corrected for the total number of electrodes pooled across subjects (N=691 electrodes). Thus, if we take a significance level of p=0.05, then the Bonferroni-corrected -log p value for significance would be 9.53. Thus, anything significant would correspond to $p<7 \times 10^{-5} (=\frac{0.05}{691})$ or, -log p>9.53, which may be approximated to 10, which corresponds to the lower limit of the color bars in the spatio-temporal plots. The activation indices for negative latencies correspond to ECoG power preceding speech onset that is correlated to speech output, while those for positive latencies correspond to ECoG power following speech onset

that is correlated to speech output. Later, this analysis was extended to twenty-six time latencies, ranging from -300 ms to 200 ms, and fifty-one time latencies, ranging from -500 ms to 500 ms, in steps of 20 ms, as discussed in Chapter 6.

Decoding Models

Following the evaluation of the spatio-temporal relationship between ECoG activity and a particular speech representation, modeling or prediction of that particular speech representation from ECoG activity was done in a subject-specific manner. For this procedure, usually ECoG activity from various time latencies are combined and used as features in the prediction model. Most of the previous studies that have tried to predict speech from ECoG activity have tried to base their predictions on either prediction of spoken language or perceived speech, but not both. One of the advantages of this dataset is the fact that it contains ECoG activity corresponding to both speech production and speech perception. Depending on what time latencies we use as the input in the prediction model, we can predict speech activity from speech production-based ECoG activity (non-positive time latencies) as well as from speech perception-based ECoG activity (non-negative time latencies). For modeling, we are interested in modeling detailed aspects of the speech activity itself, and not the intermittent periods of silence that exist in natural speech. If these silence periods are included while modeling speech, this will bias the model as the silence periods are much lower in amplitude than the speech periods. Thus, prior to modeling, the silence periods that exist in the speech activity were detected and removed. This was done by masking the speech signal by a mask that was set to 0 if speech activity

fell below a certain threshold. This threshold varied from subject to subject and the mask was verified visually to ensure that it actually masked the silence periods in the speech activity of each subject efficiently. Following the masking, the silence periods were removed from the speech representation as well as from the ECoG activity.

From the data, 70% was randomly selected and used for training the model, and the remaining 30% was used for testing the model. In the training phase of the model, 10-fold feature selection was used for the model. The ECoG feature space consists of channels $\times$ time latencies (channels $\sim$ 56-120, time latencies used for each model were usually 3, giving us a total of 168-360 features, in a subject-specific manner). Some of the finer models developed, as discussed in Chapter 6, used eight ECoG gamma sub-bands as features instead of the entire ECoG gamma band, and this increased the feature space by a factor of 8 in these models. In each fold of the feature selection, features of the ECoG activity which correlated the best with the speech representation (top 25% of ECoG features, where features are ranked in terms of r^2, where r is the correlation of a particular ECoG feature with the speech representation being modeled) were chosen. Following this 10-fold feature selection, features which were selected in at least 7 out of the 10 folds, were selected to be used in the final model. The model was then trained using the entire training data, using only the features selected, using the method of linear regression. The model was tested using 30% of the data that was set aside earlier for testing. The correlation coefficient between the predicted speech representation and the actual speech representation was used as a metric for testing the effectiveness of the model.

The statistical significance of the model for that particular speech representation was found using a bootstrapping randomization test in which the ECoG activity samples for each channel were flipped and circularly shifted by a random amount to preserve their statistical properties while destroying their temporal properties. The features which were selected using the 10-fold feature selection method as described above were chosen from this scrambled ECoG data. Following this, a model was developed using linear regression from this scrambled data to predict the particular speech representation under purview. This random circular shifting of the training data and model development were repeated 1000 times, and the squares of the correlation coefficients between the actual testing data and the testing data predicted using the models were used to form a beta distribution. The square of the actual correlation coefficient between the testing data and the data predicted using the true model was fit to the above null distribution to obtain a p-value, which gives us the statistical significance of the modeling paradigm for this particular application.

CHAPTER 4

CHARACTERIZATION AND DECODING OF PRODUCTION-BASED SPEECH REPRESENTATIONS FROM THE ELECTROCORTICOGRAM

Previous studies support the hypothesis that a strong relationship exists between ECoG activity and speech, or various aspects of speech essential for speech recognition. This chapter discusses the spatio-temporal relationships between various production-based speech representations and ECoG, as well as attempts to decode these representations directly from cortical activity. In the next chapter, similar analyses are performed for the various perception-based speech representations.

4.1 THE SPEECH POWER ENVELOPE

Characterization

As described in the preceding chapter, ECoG power was extracted in the mu, beta and the high gamma bands, to find the ECoG activity which best represents speech power in the human cortical regions corresponding to speech preparation and speech perception. Figure 8 shows the spatio-temporal correlations between ECoG high gamma power and speech power, shown for time latencies between -300 ms to 300 ms, in steps of 100 ms. When comparing Figure 8 with Figure 3 (Page 10), we see that the ECoG high gamma band power well represents the time evolution of

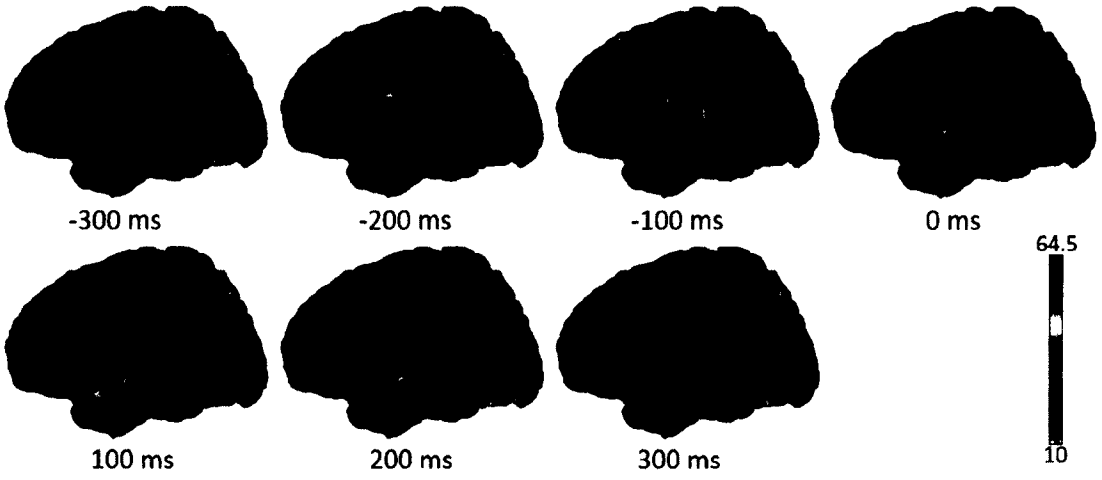


FIG. 8: Spatiotemporal correlations between the speech power and the ECoG high gamma band power, across seven time latencies relative to the onset of speech.

speech processing from preparation to perception. Activation in the negative lags is found mostly in the pre-motor and primary motor areas, and the Broca's area, which reduce in magnitude over the positive lags. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, but start at about -200 ms, strengthening at -100 ms. These spatio-temporal activations were also tested with the mu and beta band powers. However, the mu and beta powers don't represent significant activations in these speech areas. This reinforces the hypothesis that ECoG high gamma band power best represents the activation of the human cortex during speech production and processing, as was done in previous studies. Hence, we use ECoG activity in only the high gamma band for the rest of the study, and the term "ECoG high gamma band activity" is used interchangeably with the term "ECoG activity".[1]

[1]This analysis was repeated with the silence periods (between the spoken words and sentences that the subjects were silent during) removed from both ECoG activity and the speech power. The results from this analysis are shown in the Appendix.

Decoding

Using the same modeling procedure as described in Section 3.3, two models were developed for predicting the speech power, one of which was a preparation-based model (using ECoG gamma activity from time latencies -200 ms, -100 ms, 0 ms as features) and the other a perception-based model (using ECoG gamma activity from time latencies 0 ms, 100 ms, 200 ms as features). The table below shows the testing correlations for these two models, between the actual and predicted speech power, for the eight subjects.

TABLE 1: Correlation coefficients (testing) of the correlation between the actual speech and the speech predicted from the preparation-based ECoG gamma signals (combination of ECoG gamma activity at -200 ms, -100 ms, 0ms) and from the perception-based ECoG gamma signals (combination of ECoG gamma activity at 0 ms, 100 ms, 200ms). All the models were statistically significant, i.e., p-values<0.05, using the randomization test described in Section 3.3.

Sub.	A	B	C	D	E	F	G	H	Avg.
Preparation	0.04	0.14	0.18	0.13	0.09	0.07	0.09	0.20	0.12
Perception	0.06	0.11	0.17	0.12	0.12	0.11	0.07	0.18	0.12

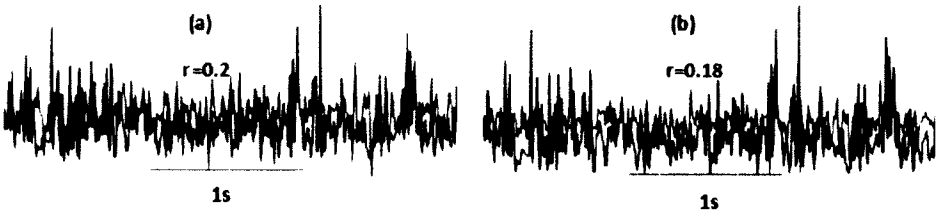


FIG. 9: Comparison of the actual (blue) and predicted (red) speech envelopes for an identical example section of the signals, as predicted by the preparation-based model (a) and the perception-based model (b). The Pearson correlation coefficient for the actual versus predicted signals, for the entire signal duration, is 0.2 and 0.18 for the preparation-based and the perception-based models respectively.

From Table 1, we can see that statistically significant models can be generated that can predict the speech power envelope moderately well from the ECoG gamma power, although the correlations are not very strong. Figure 9 depicts how the predicted speech envelopes follow the actual speech envelopes. It should also be noted that the preparatory ECoG gamma power and perceptive ECoG gamma power can almost equally well predict the speech envelope. This is a very useful revelation, because earlier studies have shown how speech can be predicted well using either preparation-based ECoG activity or perception-based ECoG activity, but not both. These results, thus, in some way, bridge the gap between preparation-based and perception-based ECoG speech studies. For the remaining speech representations in this dissertation, only preparation-based models were developed, which form the basis for an imagined speech prosthesis. However, as is shown above, both preparation-based and perception-based models can easily be developed and are almost equally good as far as predicting speech representations is concerned.

As a more in-depth analysis of the preparation-based and the perception-based models developed, the channels selected for each of the two models for each of the three time lags were studied, as shown in Figure 10. The number of features selected in each fold of the cross-validation was restricted to 20, for a more efficient visualization and interpretation. It may be observed from this figure that the channels selected for the different time latencies for the two models were varied and spread out over most of the recording areas, including the primary and secondary auditory areas which showed significant correlations with the speech power (see Figure 8). Other

studies have also shown that there is generally not a consistent cortical pattern across or within subjects for the gamma features used in speech prediction models [61].

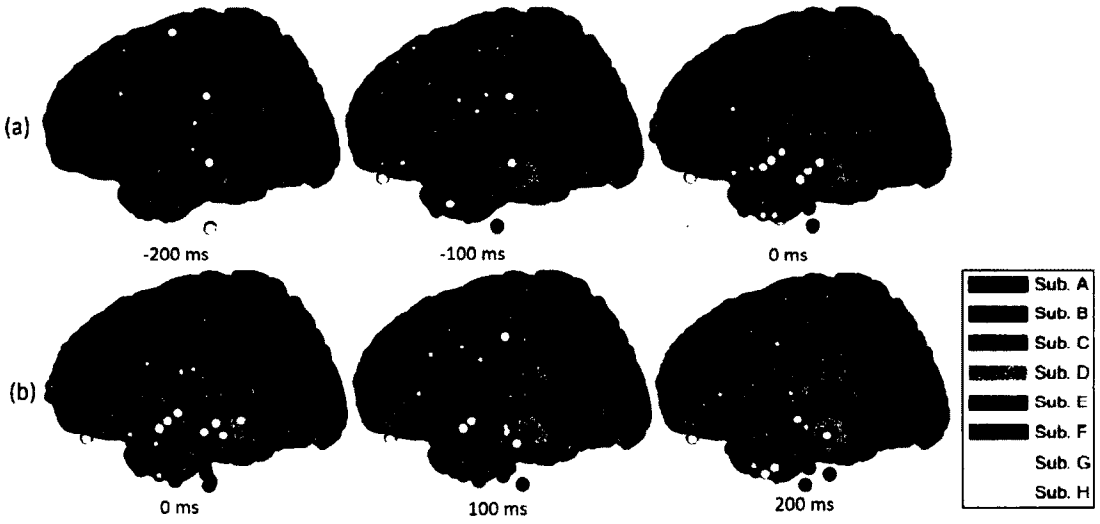


FIG. 10: Channels selected for all the eight subjects shown on a generic head model, for the preparation-based decoding model (shown in (a)) and the perception-based decoding model (shown in (b)), for the time latencies -200 ms, -100 ms, 0 ms and 0 ms, 100 ms, 200 ms respectively.

Prediction of the speech power envelope from ECoG gamma activity did not result in sufficiently high correlation coefficients for practical use. One reason for this could be the use of very simplistic feature selection and modeling techniques. Thus, in addition to using the straightforward feature selection and linear regression method as described in Section 3.3, stepwise linear regression was also implemented to determine if it can improve performance. Stepwise regression adds or removes features from a multi-linear model based on their statistical significance in a regression. Thus, it performs modeling and feature selection simultaneously to obtain the best fit in terms of the features selected and the model found over multiple iterations of the algorithm. However, it was found that the feature selection method described in

Section 3.3 in combination with linear regression works as well as, or slightly better than, stepwise linear regression, as far as modeling the speech power from ECoG gamma activity is concerned. In addition, instead of using linear regression to solve for the model after feature selection, a basic neural network with one hidden layer was trained using the conjugate gradient method with back-propagation, to predict speech power from ECoG gamma activity. A more optimized set of features were also developed for predicting the speech power, which led to improved predictions. The results from this optimized feature set and improved decoding models are discussed in Chapter 6.

4.2 THE FUNDAMENTAL FREQUENCY AND FORMANTS

Acoustic speech waves demonstrate the sounds radiated as pressure modulations from the lips while articulating language. The amplitude as well as the frequency of this speech wave varies with time. A speech waveform typically consists of two parts: (a) a quasi-periodic part, which is repetitive or periodic over a brief period of time and (b) a noise-like part, of random amplitude and frequency. The frequency of the quasi-periodic part is called the fundamental frequency or the pitch frequency, denoted by F0. F0 usually varies slowly over time for a speech signal. F0 is one of the most important acoustic features for phoneme identification, and is used in low bit rate systems for reconstructing speech. F0 can be 80 Hz or lower for male adults and above 300 Hz for female adults and children. Other than the fundamental frequency, spectral characterization of speech is also often performed using formants. Formants are the spectral peaks which occur in the speech spectrum [74] which also represent

the resonances of the human vocal tract. Formants are often used for speech recognition because the information that humans require to distinguish between vowels is represented purely in the frequency content of the vowel sounds. The formant with the lowest frequency is called F1, the second F2, the third F3. Most often, the first two formants, F1 and F2 are enough to distinguish between vowels. Using the overall frequency spectrum of speech, it was found that the speech power is mostly concentrated at the lower frequencies, i.e., over 80% of speech power lies below 1 kHz. The normal range of the formant frequencies for adult males is F1=180-800 Hz, F2=600-2500 Hz, F3=1200-3500 Hz. For adult females, the formant frequencies are about 20% higher than those for adult males [69]. Separate time and frequency analyses of speech sometimes do not reflect important properties of the speech signal, such as the frequency content of the speech signal varying continuously over time, which is why we need a joint time-frequency representation of speech. The speech spectrogram is a commonly used time-frequency representation of speech. It is a three-dimensional representation of speech, where the short-time Fourier transform is plotted with time units on the horizontal axis, frequency is plotted on the vertical axis, and power is represented in the third dimension, with different colors or levels of darkness showing the different power levels. Spectrogram reading, i.e., the manual process of using visual displays of the spectrogram to identify and decode semantic information, is often undertaken to understand spoken language from the simultaneous availability of temporal and spectral information in the spectrogram [69]. Formants can be measured as the amplitude peaks in the frequency spectrum

of speech, using the speech spectrogram.

Using the original speech signals recorded for the eight subjects in this study, the fundamental frequency and the first two formants were extracted. This was done using the Praat software [87], a popularly used software for speech analysis and synthesis. This software uses spectral analysis to compute F0, F1 and F2 at a rate of 100 Hz. Thus, 96 values for F0, F1 and F2 were computed per second, since the original sampling rate of the speech signal was 9600. F0, F1 and F2 vary almost instantaneously with time, as does the speech amplitude. The values for F0, F1 and F2 computed are approximations of the actual values, although the computations are done at a high enough sampling rate for the approximations to be accurate enough.

Characterization

The spatio-temporal correlations between the fundamental frequency, first formant, second formant and the ECoG high gamma power was computed in the same manner as described in Section 3.3. For the first two formants, the spatio-temporal correlations achieved were not statistically signficant for any of the channels and hence are not shown here. The spatio-temporal correlations between the fundamental frequency and the high gamma power are shown in Figure 11. When comparing this figure with Figure 3 (Page 10), we see that similar regions are activated for the fundamental frequency as were activated for the speech power, i.e., activation in the negative lags is found mostly in the pre-motor areas, the Broca's area and the primary motor areas, which reduce in magnitude over the positive lags. Activation in the superior temporal gyrus and the Wernicke's area is found mostly at the positive

lags, but start at about -200 ms, strengthening at -100 ms.[2]

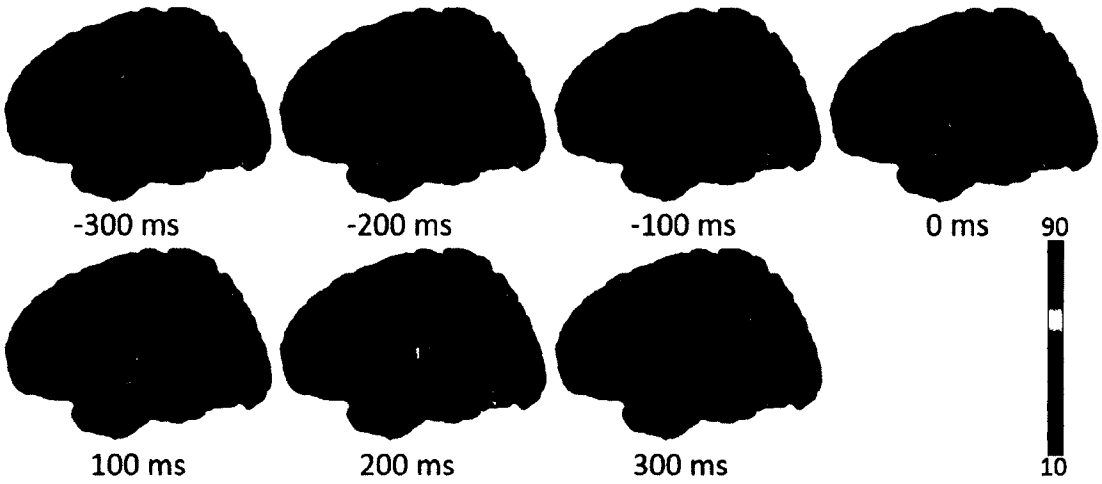


FIG. 11: Spatiotemporal correlations between ECoG high gamma power and the fundamental frequency, across seven time latencies relative to the onset of speech.

Decoding

Using the same modeling procedure as described in Section 3.3, preparation-based models (using ECoG gamma activity from time latencies -200 ms, -100 ms, and 0 ms as features) were developed for predicting the fundamental frequency and the first two formants directly from ECoG gamma activity. The table below shows the testing correlations for these models, between the actual and predicted fundamental frequency and the first two formants, for the eight subjects.

[2]This analysis was repeated with the silence periods (between the spoken words and sentences that the subjects were silent during) removed from both ECoG activity and the fundamental frequency. The results from this analysis are shown in the Appendix.

TABLE 2: Correlation coefficients (testing) of the correlation between the actual and the predicted fundamental frequency or F0 (first row), first formant or F1 (second row), second formant or F2 (third row) from the preparation-based ECoG gamma power (combination of ECoG gamma activity at -200 ms, -100 ms, 0ms). Some of the models, marked by a superscript 'x' were not statistically significant, i.e., p-values>0.05, using the randomization test described in Section 3.3. The rest of the models were statistically significant, i.e., p-value<0.05.

Sub.	A	B	C	D	E	F	G	H	Avg.
F0	0.08	0.2	0.09	0.09	0.03^X	0.12	0.08	0.34	0.13
F1	0.08	0.23	0.03^X	0.01^X	0.02^X	0.11	0.02^X	0.14	0.08
F2	0.08	0.14	0.08	0.03^X	0.06	0.1	0.13	0.12	0.09

From Table 2, it is observed that these models provide some degree of prediction of the fundamental frequency and the first two formants from the ECoG gamma power. The fundamental frequency and the second formant can be modeled relatively well for most of the subjects (statistically significant correlations for seven out of eight subjects), although the correlation coefficients are not as high as would be practically useful. The models used to predict the first formant are much worse in terms of statistical significance. In general, however, most of these models are not accurate enough to be used in practice. Thus, for practically useful predictions, the models need to be developed and optimized further to give better prediction performance. In this direction, a more optimized model was developed for predicting the fundamental frequency from ECoG gamma activity, using a more optimized set of features and a simple neural network, which led to improved predictions. The results from this optimized model are discussed in Chapter 6.

4.3 LINEAR PREDICTIVE CODING

Linear Predictive Coding (LPC) is another popular speech analysis method [88], in which linear prediction is used to linearly combine the past time-domain samples, s[n-1], s[n-2],..., s[n-M] to predict the present time-domain sample s[n], as follows:

$$s[n] \approx \hat{s}[n] = -\sum_{i=1}^{M} a_i s[n-1] \tag{3}$$

where a_i, i =1, 2, ... M are called the LPC coefficients and ŝ [n] is the predicted current time sample. The LPC coefficients can be taken directly from the speech signal s[n] by minimizing the square of the error signal,

$$e[n] = s[n] - \hat{s}[n] = s[n] + \sum_{i=1}^{M} a_i s[n-1] = \sum_{i=0}^{M} a_i s[n-1] \tag{4}$$

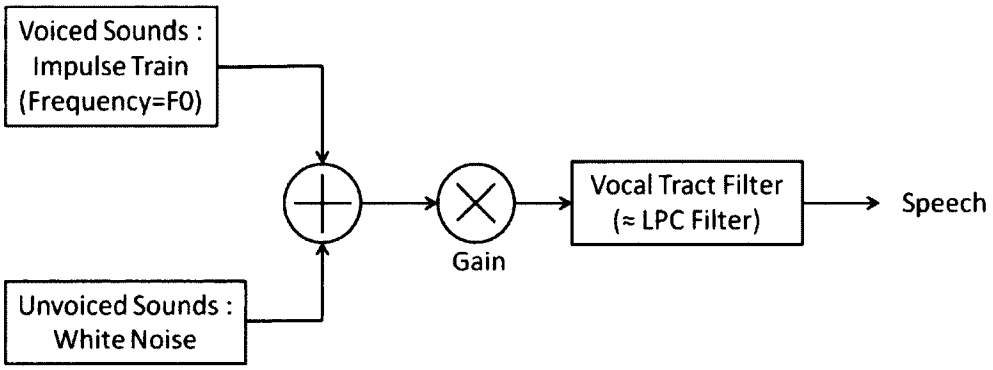


FIG. 12: The source-filter model of speech production

This least squares optimization can be done using the autocorrelation method or the covariance method. One of the most important applications of LPC is in the

source-filter model of speech production [69]. The source-filter model of speech production models speech as a combination of a sound source such as the vocal cords and the vocal tract filter. In the implementation of the source-filter model, the sound source is modeled as a periodic pulse train, the period being the inverse of the fundamental frequency, F0, for voiced speech and as white noise for unvoiced speech. The time-varying vocal tract filter is modeled simplistically as an all-pole filter, the coefficients of which are the LPC coefficients. The order of, i.e., the number of poles in the all-pole LPC model, should be chosen so as to obtain the optimal trade-off between the level of detail (higher the order, higher the detail) that should be incorporated and the complexity of computation (lower the order, lower the complexity). Using the original speech signals recorded in this study, the linear prediction coefficients were estimated using MATLAB. This function determines the coefficients of a forward linear predictor by minimizing the prediction error using a least squares cost function. A 10^{th} order linear predictor, i.e., FIR filter was used that predicts the current speech sample value based on 10 past samples, using autoregressive modeling. This was done for every 100 samples of the speech signal, i.e., one set of LPC coefficients were found for every 96 ms of speech. Thus, the LPC coefficients were extracted at a sampling rate of 100 Hz.

Characterization

The spatio-temporal correlations between the LPC coefficients and the ECoG high gamma power was computed in the same manner as described in Section 3.3. The correlations obtained were not statistically significant for all of the LPC coefficients.

The spatio-temporal correlations for the significant LPC coefficients are as shown in Figure 13.

For all the significant LPC coefficients, activations were found in the pre-motor and motor cortex areas prior to speech onset. The activations in the auditory areas, i.e., the superior temporal gyrus, Broca's area and Wernicke's area, started slightly before speech onset and continued until after speech onset. This means the LPC coefficients, which are also equivalent to, the vocal tract filter coefficients for speech, activates similar regions in the human cortex as does the speech power and the fundamental frequency or pitch of speech. It is also interesting to note that some filter coefficients significantly activate the speech areas, while other coefficients do not. Also, among the significant coefficients, some coefficients demonstrate stronger activations than others do. For example, coefficient numbers 2 and 4 demonstrate the strongest activations here. This may mean that even numbered past samples hold the most information as far as prediction of the current speech sample is concerned. This may have implications for predicting speech from LPC coefficients. It can be concluded that predicting the 2^{nd} and 4^{th} LPC coefficients accurately is vital for speech. [3]

[3]This analysis was repeated with the silence periods (between the spoken words and sentences that the subjects were silent during) removed from both ECoG activity and the LPC coefficients. The results from this analysis are shown in the Appendix.

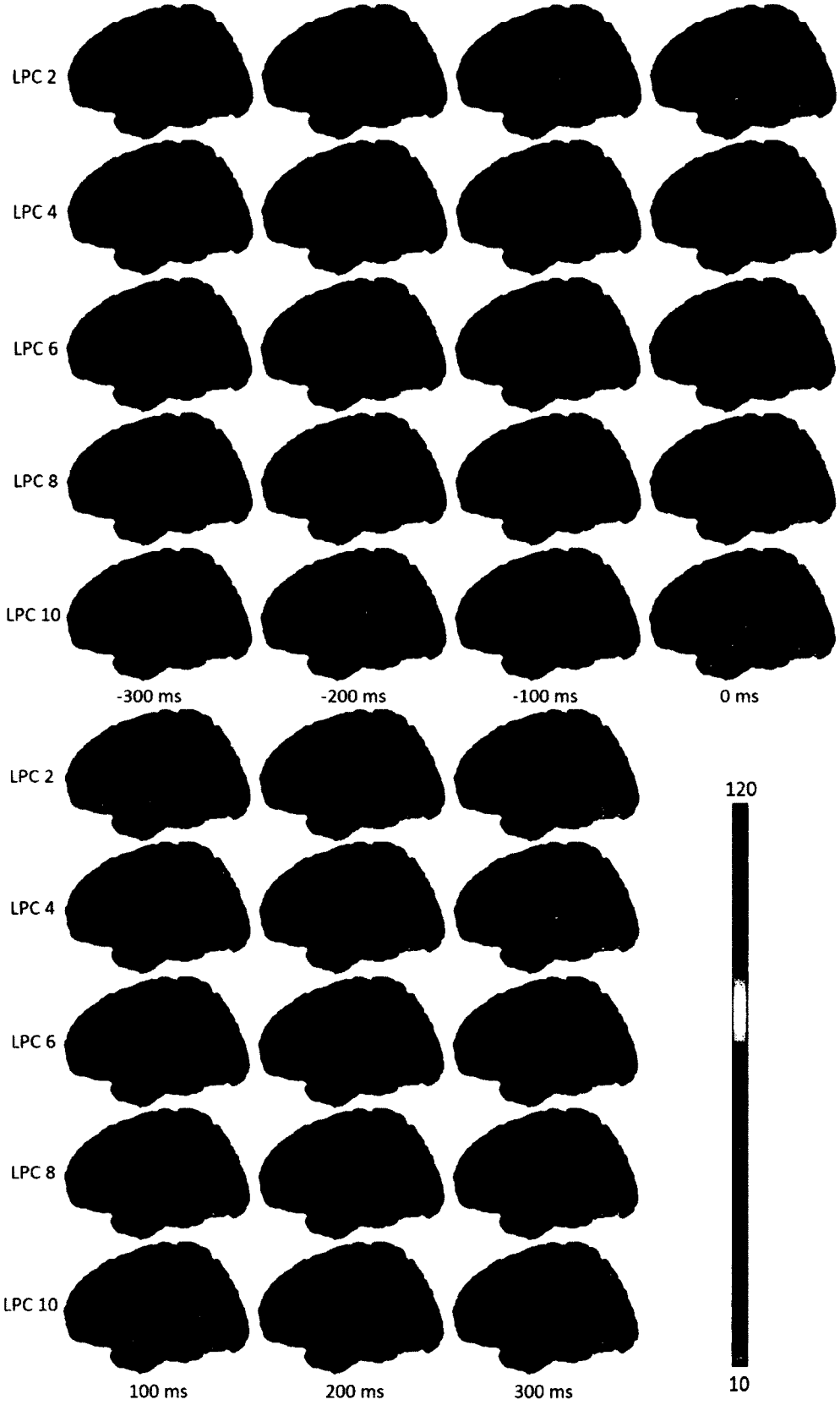


FIG. 13: Spatiotemporal correlations between ECoG high gamma power and the significant LPC coefficients (shown in the five rows respectively) across seven time latencies relative to the onset of speech.

Decoding

Using the same modeling procedure as described in Section 3.3, preparation-based models (using ECoG gamma activity from time latencies -200 ms, -100 ms, and 0 ms as features) were developed for predicting the ten LPC coefficients directly from ECoG gamma activity. Table 3 shows the testing correlations between the actual and predicted 10 LPC coefficients, averaged over the eight subjects.

TABLE 3: Average of the correlation coefficients (testing) of the correlation between the ten actual and the predicted LPC coefficients (shown in the ten columns, numbered 1-10, respectively)from the preparation-based ECoG gamma power (combination of ECoG gamma activity at -200 ms, -100 ms, 0ms). The number shown in parantheses is the number of subjects for which the correlations were statistically significant, i.e., $p<0.05$, for that particular LPC coefficient.

	1	2	3	4	5	6	7	8	9	10
Avg.	0.13	0.12	0.1	0.08	0.08	0.07	0.07	0.06	0.06	0.08
Corr.	(8)	(8)	(7)	(7)	(6)	(8)	(7)	(6)	(6)	(8)

The first two LPC coefficients could be predicted the best among all the coefficients. This makes sense because the LPC coefficients represent an auto-regressive modeling of the speech signal. Thus, the lower coefficients contain the most information as far as the present speech sample is concerned. However, the remaining coefficients are also important as the more the coefficients that can be predicted from ECoG activity, the better speech can be reconstructed. For this reason, it is important that the LPC coefficients be predicted to a better extent from ECoG high gamma activity than is obtained here. This will possibly require more sophisticated

and optimized modeling procedures which will be able to better model the relationship between ECoG high gamma power and the LPC coefficients. This will be useful so that we can then use the ECoG-predicted LPC coefficients to reconstruct the speech activity directly.

4.4 CONCLUSION

The production-based speech representations discussed here, namely, the speech power, fundamental frequency, formants, and the linear predictive coefficients activate similar regions of the human motor and auditory cortex as shown in previous studies. Some of these features can be predicted relatively well from ECoG high gamma power while others cannot be predicted significantly well. This can be attributed to a variety of factors including suboptimal electrode locations for given representations, the nature of a given speech representation in the cortex, etc. However, these results show that the cortical activity does contain information relevant to these representations and that can be used for speech prediction. An area of prospective work would be to develop more optimized models which can predict these representations more accurately than the models discussed here. Another area of future research would be to use a combination of these predicted production-based speech reconstructions to attempt various speech recognition tasks, so as to develop an ECoG-based automatic speech recognition system, or to use these ECoG decoded production-based speech representations to reconstruct the speech signal itself.

CHAPTER 5

CHARACTERIZATION AND DECODING OF PERCEPTION-BASED SPEECH REPRESENTATIONS FROM THE ELECTROCORTICOGRAM

The previous chapter shows that strong spatio-temporal relationships exist between ECoG activity and various production-based speech representations such as speech power, fundamental frequency, formants and linear predictive coefficients of speech, and these speech representations can be decoded to some extent directly from ECoG activity. Here, the spatio-temporal relationships between various perception-based speech representations and ECoG are discussed, as well as approaches to decode these representations directly from cortical activity.

5.1 THE MEL FREQUENCY CEPSTRUM

The speech signal is generally considered as the convolution of an excitation waveform or a sound source waveform with the vocal-tract filter response [69]. Some speech applications require separate estimation of these two components, i.e., a de-convolution of the excitation and filter response components is beneficial. This is done by the method of cepstral analysis or cepstral de-convolution [89]. Cepstral de-convolution converts the product of two spectra into a sum of two signals, which may further be separated by linear filtering if they are different enough. If we assume

that the speech spectrum, S, is equal to the product of an excitation signal, E, and the vocal-tract spectrum, H, then cepstral de-convolution can be used to linearly separate these two components. This is typically done by taking the logarithm of the speech spectrum, which gives us the following relationship:

$$logS = log(EH) = log(E) + log(H) \tag{5}$$

Since H consists of smooth formant transitions, i.e., a slowly varying frequency spectrum, while E is more irregular, due to noise or impulse train excitation, these two components can be linearly separated with ease. Cepstral analysis is often combined with a non-linear weighting in the frequency domain of the speech spectrum. The mel-frequency cepstral coefficients (MFCCs) are obtained by mapping the speech spectrum onto the mel scale, which is a non-linear logarithmic scale, based on the perception of pitch by the human ear. The mel scale is related to the frequency scale as follows [90]:

$$m(\omega) = 1127\ ln(1 + \frac{\omega}{1400\pi}) \tag{6}$$

Following this, the logarithm of this warped power spectrum is taken, in accordance with the cepstral analysis methodology. Finally, the inverse Discrete Fourier Transform or the Discrete Cosine Transform is used to obtain the MFCCs. The first M coefficients (M $\sim$ 8-14), leaving aside the initial coefficient (the initial coefficient usually just represents the average speech power) then represent the MFCCs. Using the original speech signals recorded in this study, the mel frequency cepstral coefficients were estimated using the RASTA-PLP toolbox in MATLAB. The speech signal was

divided into overlapping frames of length 256 samples, each of which was then windowed by multiplying it with the Hamming window. The power of the obtained spectrum was mapped onto the mel scale, the transformation that accounts for non-linear pitch perception. The logarithm of the power was then computed from the spectrum, and the discrete cosine transform was taken from 12 mel frequency bands to obtain the twelve MFCCs.

Characterization

The spatio-temporal correlations between the MFC coefficients and the ECoG high gamma power was computed in the same manner as described in Section 3.3. The correlations obtained were statistically significant only for the first three MFC coefficients, as shown in Figure 14. For these MFC coefficients, activations in the pre-motor and motor cortex areas are found prior to speech onset. Activations in the auditory areas, i.e., the superior temporal gyrus, Broca's area and Wernicke's area, start slightly before speech onset and continue until after speech onset. Thus, the MFC coefficients, which represent a perceptual pitch-based transformation of speech, activate similar regions in the human cortex as do the speech power and the fundamental frequency (pitch) of speech. This is somewhat expected because the MFC coefficients are representative of a pitch-based warping of the original speech spectrum. Thus, the MFC coefficients essentially contain the power information and the pitch information of a speech signal. It is also interesting to note that only the lowest three MFC coefficients significantly activate the speech areas, while the other coefficients do not. The different MFCC coefficients correspond to different

filters arranged in decreasing order of pitch perception. The coefficients with the significant activations correspond to mel filters with center frequencies which are the most important for the human perception of pitch.[1]

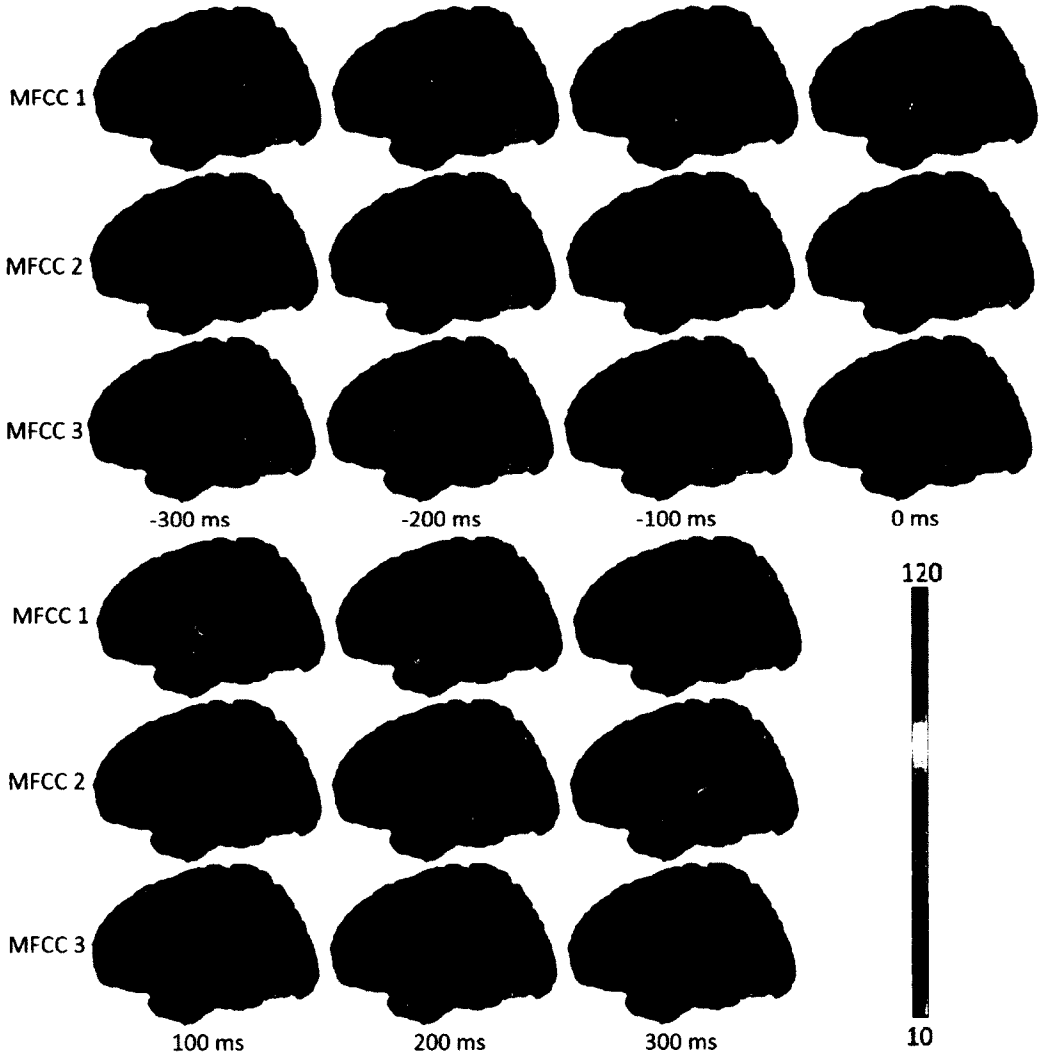


FIG. 14: Spatiotemporal correlations between ECoG high gamma power and the significant MFC coefficients (shown in the three rows respectively) across seven time latencies relative to the onset of speech.

[1]This analysis was repeated with the silence periods (between the spoken words and sentences that the subjects were silent during) removed from both ECoG activity and the MFC coefficients. The results from this analysis are shown in the Appendix.

Decoding

Using the same modeling procedure as described in Section 3.3, preparation-based models (using ECoG gamma activity from time latencies -200 ms, -100 ms, and 0 ms as features) were developed for predicting the twelve MFC coefficients directly from ECoG gamma activity. The RASTA-PLP toolbox which was used to extract the MFCCs may also be used to invert the extraction process, to get back the speech signal from the MFCCs. Thus, if the MFCCs can be predicted well enough from ECoG high gamma power, these decoded MFCCs can then be used to get back the original speech signal. Table 4 shows the testing correlations between the actual and predicted twelve MFC coefficients, averaged over the eight subjects.

TABLE 4: Average of the correlation coefficients (testing) of the correlation between the actual and the predicted 12 MFC coefficients (shown in the twelve columns, numbered 1-12, respectively) from the preparation-based ECoG gamma power (combination of ECoG gamma activity at -200 ms, -100 ms, 0ms). The number shown in parantheses is the number of subjects for which the correlations were statistically significant, i.e., $p<0.05$, for that particular MFC coefficient.

	1	2	3	4	5	6	7	8	9	10	11	12
Avg.	0.12	0.16	0.16	0.13	0.13	0.12	0.08	0.08	0.1	0.1	0.09	0.1
Corr.	(8)	(8)	(8)	(8)	(8)	(7)	(7)	(7)	(7)	(7)	(7)	(7)

The first five MFC coefficients could be predicted the best among all the coefficients. For some subjects, such as Subject B and Subject H, all the MFCCs could be predicted moderately well. For all the subjects except Subject E, the predictions of all the MFCCs were found to be statistically significant. In general, the MFCCs could be predicted better than the LPC coefficients. Thus, these decoded MFCCs

could potentially be used to reconstruct the original speech signal. However, before doing so, it is important to further improve the prediction correlations. This will likely require more sophisticated and optimized modeling procedures that will be able to better model the relationship between ECoG high gamma power and the MFC coefficients.

5.2 PERCEPTUAL LINEAR PREDICTION

A variation of the basic linear prediction method is the Perceptual Linear Prediction (PLP), wherein the power spectrum of the speech signal is modified prior to its approximation by the autoregressive model. This modification is done by first mapping the power spectrum of speech onto the Bark scale, which is a scale based on the perception of loudness by the human ear. The Bark scale is related to the frequency scale as follows:

$$\Omega(\omega) = 6\ ln(\frac{\omega}{1200\pi} + \sqrt{(\frac{\omega}{1200\pi})^2 + 1}) \tag{7}$$

where ω is the angular frequency in rad/s [79]. This scale models the ear's behavior in perceiving loudness as a function of frequency. Following this mapping, the resulting power spectrum is convolved with the power spectrum of the critical-band

masking curve $\psi(\Omega)$.The critical-band curve is a piece-wise function given by:

$$\begin{aligned}
\psi(\Omega) &= 0 && if\ \Omega < 1.3 \\
&= 10^{2.5(\Omega+0.5)} && if\ -1.3 < \Omega < -0.5 \\
&= 1 && if\ -0.5 < \Omega < 0.5 \\
&= 10^{-1.0(\Omega-0.5)} && if\ 0.5 < \Omega < 2.5 \\
&= 0 && if\ \Omega > 2.5
\end{aligned} \tag{8}$$

This piece-wise function is a crude approximation to the shape of the human auditory filters. The convolution of $\psi(\Omega)$ with the Bark scale mapped power spectrum, P(Ω) gives the samples of the critical-band power spectrum, as follows:

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\psi(\Omega) \tag{9}$$

The speech spectrum is then re-sampled at intervals of 1 Bark, followed by which an equal loudness curve is used to pre-emphasize it. The equal-loudness function is an approximation of the non-equal sensitivity of the human hearing system to different frequencies. A popular equal-loudness function curve is given by:

$$E(\omega) = \frac{[(\omega^2 + 56.8 * 10^6)\omega^4]}{(\omega^2 + 6.3 * 10^6)^2(\omega^2 + 0.38 * 10^9)} \tag{10}$$

adopted from Makhoul and Cosell [79]. The last operation prior to the autoregressive modeling is where the cubic root of the above spectrum is taken. This is known as the cubic-root amplitude compression, and is an approximation of the power law of hearing and models the non-linear relationship between the amplitude of sound and its perceived loudness [79]. In the final step, the spectrum resulting from all the above steps is approximated by an all-pole model, as is done in LPC coding. The

resulting coefficients are called the PLP coefficients. Using the original speech signals recorded in this study, the perceptual linear prediction coefficients were estimated using the RASTA-PLP toolbox in MATLAB. The power spectrum was computed and grouped into critical bands, the transformation that accounts for non-linear loudness perception. The logarithm of the power was then computed from the spectrum, followed by which auditory cubic compression was performed. Finally, the linear prediction analysis was done on this spectrum with an auto-regressive model order of 10 to obtain the 10 PLP coefficients, after ignoring the 0th coefficient as it is equivalent to the speech power.

Characterization

The spatio-temporal correlations between the PLP coefficients and the ECoG high gamma power was computed in the same manner as described in Section 3.3. These correlations were found to be statistically significant only for the first PLP coefficient, shown in Figure 15. For this PLP coefficient, activations in the pre-motor and motor cortex areas were found prior to speech onset and activations in the auditory areas, i.e., the superior temporal gyrus, Broca's area and Wernicke's area, were found to start slightly before speech onset, continuing until after speech onset. This means that PLP, which is derived from a loudness-based linear predictive coding representation of speech, activates similar regions in the human cortex as do the speech power and the fundamental frequency of speech. The different PLP coefficients correspond to different filters arranged in decreasing order of loudness perception. Since only the first coefficient was found to have significant activations, this corresponds to the

Bark filter with a center frequency most important for loudness perception, and to the autoregressive model coefficient which holds the most amount of information for predicting the current speech sample.[2]

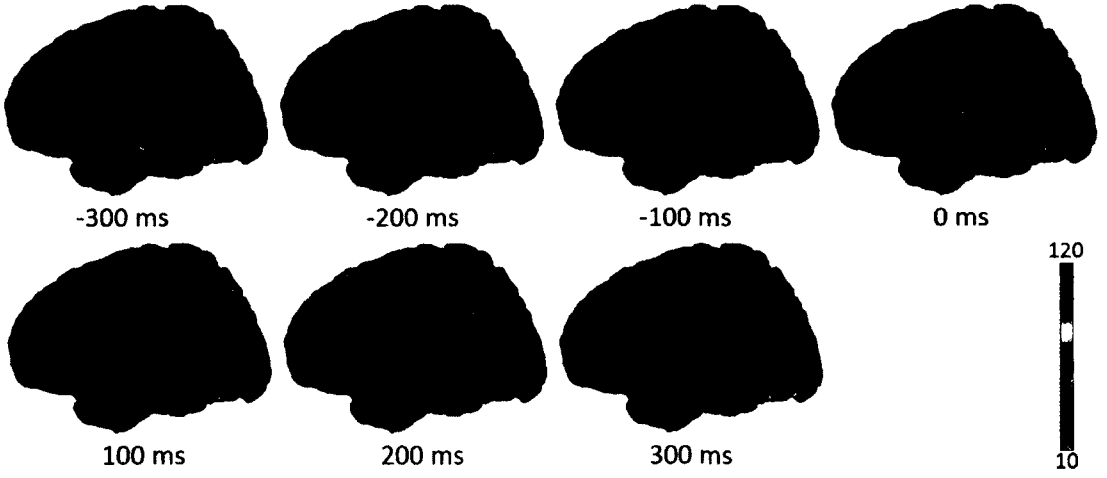


FIG. 15: Spatiotemporal correlations between ECoG high gamma power and the first PLP coefficient across seven time latencies relative to the onset of speech.

Decoding

Using the same modeling procedure as described in Section 3.3, preparation-based models (using ECoG gamma activity from time latencies -200 ms, -100 ms, and 0 ms as features) were developed for predicting the ten PLP coefficients directly from ECoG gamma activity. The RASTA-PLP toolbox which was used to extract the PLP coefficients may also be used to invert the extraction process, in order to recover the speech signal from the coefficients. Thus, if the PLP coefficients can be predicted well enough from ECoG high gamma power, they can be used to reconstruct the

[2]This analysis was repeated with the silence periods (between the spoken words and sentences that the subjects were silent during) removed from both ECoG activity and the PLP coefficients. The results from this analysis are shown in the Appendix.

original speech signal. Table 5 shows the testing correlations between the actual and predicted 10 PLP coefficients, averaged over the eight subjects.

TABLE 5: Average of the correlation coefficients (testing) of the correlation between the ten actual and the predicted PLP coefficients (shown in the ten columns, numbered 1-10, respectively) from the preparation-based ECoG gamma power (combination of ECoG gamma activity at -200 ms, -100 ms, 0ms). The number shown in parantheses is the number of subjects for which the correlations were statistically significant, i.e., $p<0.05$, for that particular PLP coefficient.

	1	2	3	4	5	6	7	8	9	10
Avg.	0.12	0.17	0.13	0.13	0.14	0.12	0.12	0.11	0.1	0.11
Corr.	(7)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)

All the PLP coefficients could be predicted relatively well, as compared to the MFCC and the LPC representations of speech. Thus, warping the speech spectrum onto the loudness-based Bark scale, improves the prediction of the linear prediction coefficients. These decoded PLP coefficients could potentially be used to reconstruct the original speech signal. This would, however, require further improvements in the prediction correlations. This will require more sophisticated modeling procedures which can better model PLP coefficients from ECoG high gamma power.

5.3 CONCLUSION

The perception-based speech representations discussed here activate similar regions of the human motor and auditory cortex as shown in previous studies. The models used to predict these speech features should be optimized to improve the prediction so it becomes practical for speech reconstruction, or for use in cortically driven automatic speech recognition systems.

CHAPTER 6

OPTIMIZATION OF THE CHARACTERIZATION AND THE DECODING MODELS

The characterization and decoding models discussed in the previous two chapters, in particular, the decoding models, showed promise but were clearly suboptimal. This chapter discusses the development of more optimized spatio-temporal characterization models and improved decoding models, for the characterization of speech power and the decoding of speech power and the fundamental frequency (pitch information) from ECoG gamma activity.

6.1 OPTIMIZED CHARACTERIZATION

The characterization techniques used in the previous chapters for computing spatio-temporal correlations and their significance may be further optimized to determine finer details in the relationships between ECoG activity and speech and its representations. Two possible directions are discussed, which lead to more optimized characterization of speech activity in the human cortex.

6.1.1 HIGHER TEMPORAL RESOLUTION

The spatio-temporal correlations that were plotted for the previous chapters were at a time resolution of 100 ms. This is a relatively large time resolution, and it would

be of interest to determine if smaller changes in cortical activity can be detected, relative to speech progression. The advantage of using ECoG becomes evident with such an analysis as, fortunately, very high temporal resolutions are available with ECoG as compared to other hemodynamic signal acquisition techniques such as fMRI and PET. Thus, a time resolution of 20 ms was used to investigate the spatio-temporal correlations of ECoG high gamma activity relative to speech activity, starting from 300 ms prior to speech onset, and continuing up to 200 ms post speech onset.

To further evaluate the contribution of the different speech areas of interest in the cortex, the electrode locations for all the subjects were classified into the seven areas of interest for the purpose of speech planning, production and articulation, and perception, namely the Broca's area, the premotor cortex, motor cortex, the mSTG (middle Superior Temporal Gyrus), the pSTG (posterior Superior Temporal Gyrus), the MTG (Middle Temporal Gyrus) and the SMG (SupraMarginal Gyrus). These areas are shown in Figure 3 (Page 10). This classification was done in two steps. First, the electrode locations for all the subjects were labeled as the Brodman areas they belonged to [91]. Then, these regions were classified as either belonging to any of the seven cortical regions of interest, or belonging to none of the seven regions, using a combination of the knowledge of the Brodmann area they belonged to, as well as by a manual thresholding procedure primarily based on visualization. The resulting spatio-temporal correlations are shown in Figure 16.

As may be expected, a more gradual progression in activations are evident in this plot, as compared to Figure 8 (Page 48). It is also interesting to note that the

Broca's area, most involved in speech planning and articulation, starts to become activated at -300 ms, and remains strongly activated until about 20 ms (a small time lag after actual speech onset), and concludes after 80 ms. The activations of the pre-motor cortex, also involved in speech planning, also start at -300 ms and slowly reduce and evolve into activations of the motor cortex, which is involved in mouth movements related to speech articulation, around -40 ms (a small time lead prior to speech onset). The activations of the Wernicke's area and the superior temporal gyrus, most involved in speech perception, begin to show significantly around -120 ms and continue getting stronger until 100 ms, after which they start to reduce as shown in Figure 16. Although the advantage of using a smaller time resolution is not that evident in this particular figure, it may be useful for detecting small temporal changes in certain cortical regions in spatio-temporal activation plots, which can show drastic, sudden changes when using higher time resolutions such as 100 ms.

Figure 17 quantitatively summarizes the activation indices as averages over the significant electrodes in each of the seven regions, over different time latencies, while Figure 18 demonstrates the time progression of the average of the significant activations in each of the seven cortical regions of interest. Each row in these figures correspond to one relevant cortical area, and the rows are topographically ordered, from the anterior (or frontal) areas to the posterior (or temporal) areas. Each column in Figure 17 and each time point in Figure 18 represents a time latency from -500 ms to 500 ms, with time steps of 20 ms. It should be noted that a larger time range was chosen for these two figures, as compared to the usual -300 to 300 ms chosen for most

of the earlier spatio-temporal activation figures. This is because in the earlier figures, some areas were significantly activated even at -300 ms or at 300 ms, which raises the question as to whether different speech areas are activated beyond this range of time latencies. It is observed from these figures that activations in some speech areas start as early as -500 ms, although the most significant activations for most regions are observed after -300 ms. Similarly, activations in certain speech regions continue until 400 ms, and gradually reduce. These quantitative analyses also demonstrate that the frontal regions show stronger and more significant (more red colors in Figure 17 and higher values in Figure 18) activations at negative lags, while the temporal regions demonstrate more significant and stronger activations at positive time latencies. Activation indices for the supramarginal gyrus (which was primarily a control location for this analysis) are negligible for all the time latencies.

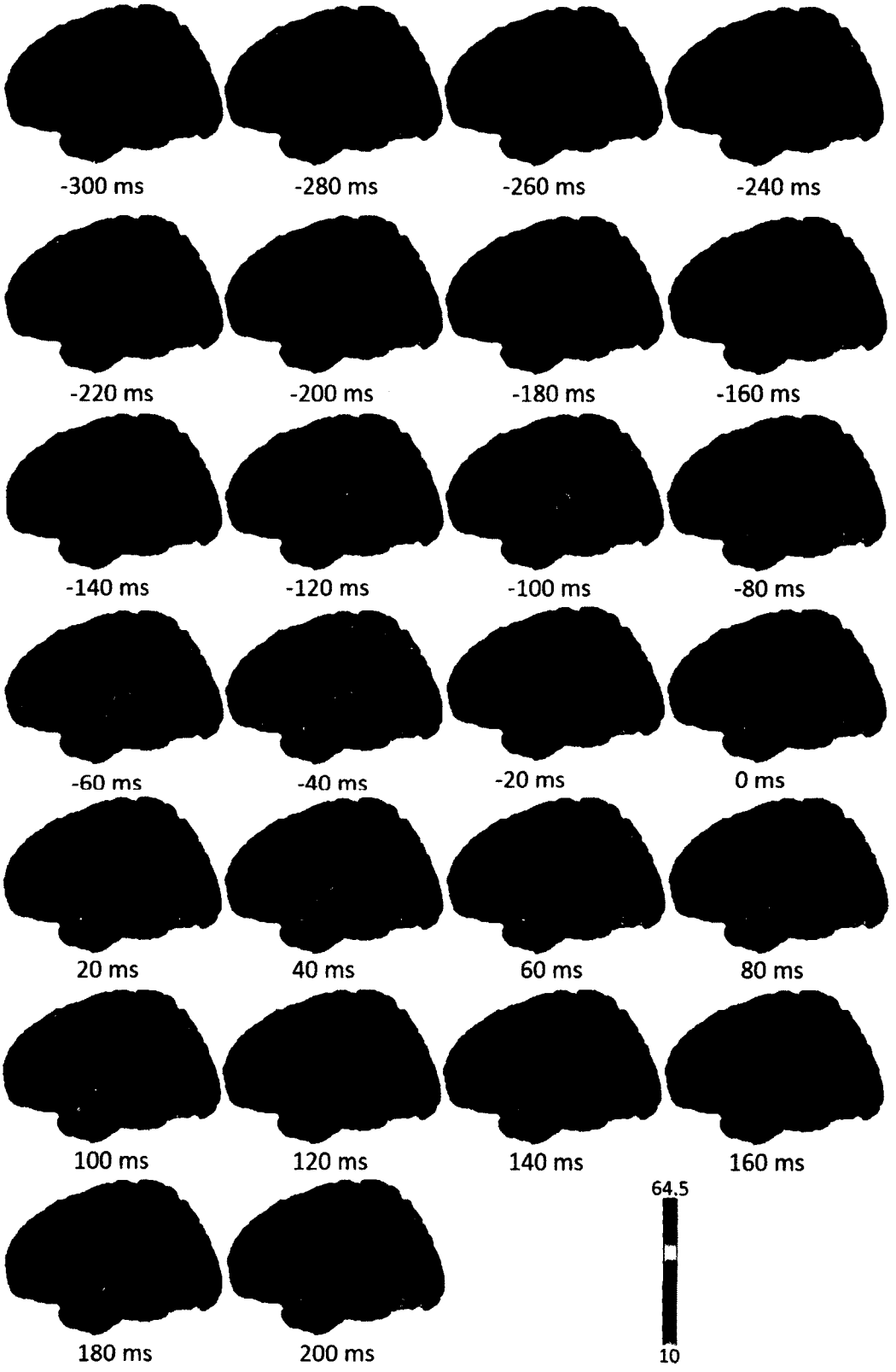


FIG. 16: Spatiotemporal correlations between ECoG high gamma power and the speech power envelope, across time latencies ranging from -300 ms to 200 ms, relative to the onset of speech, in steps of 20 ms [92].

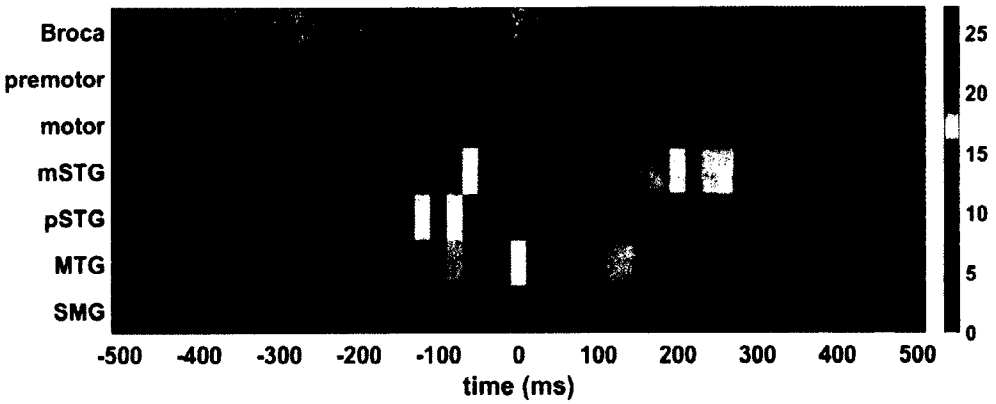


FIG. 17: Average activation indices for electrodes with statistically significant correlations in the seven cortical regions of interest, shown across time latencies from -500 ms to 500 ms. The areas of interest include the three frontal areas (Broca's area, premotor and primary motor cortex), three temporal areas (middle and posterior superior temporal gyrus [mSTG and pSTG respectively] and the middle temporal gyrus [MTG]) and one parietal area [supramarginal gyrus]. The activation index is indicated by a color scale, where in areas with less statistically significant correlations with the speech power are represented with more blue colors while areas with larger statistical significance of correlations are represented with more red colors [92].

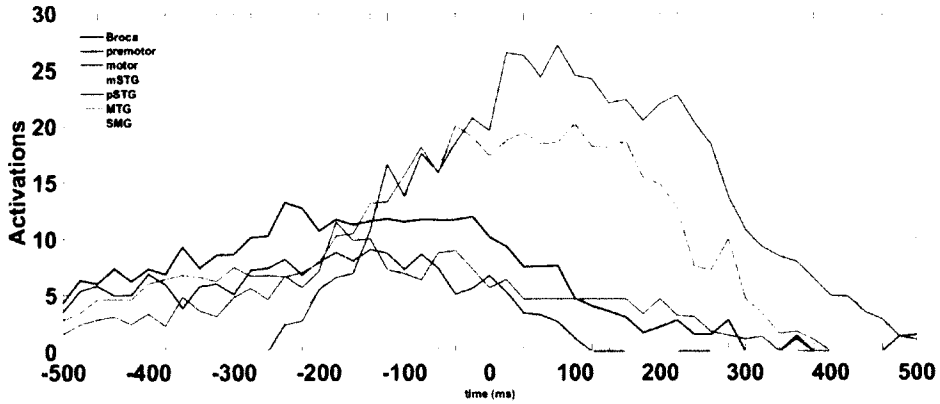
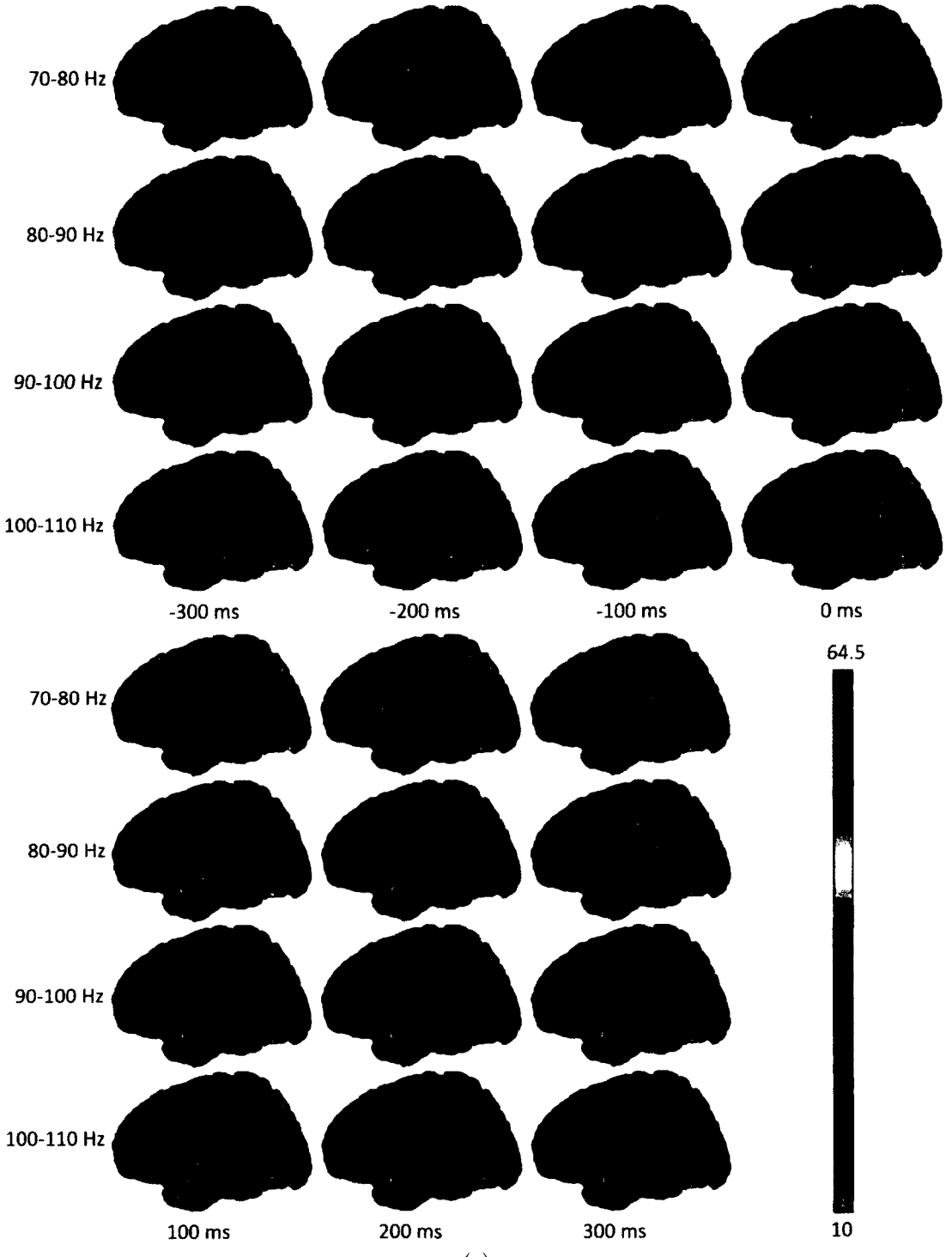


FIG. 18: Time progression of the average activation indices of the seven cortical areas of interest, as a function of time. This is another representation of Figure 17, with a better demonstration of the temporal progression of the statistical significance of correlations of the seven different cortical regions of interest with the speech power [92].

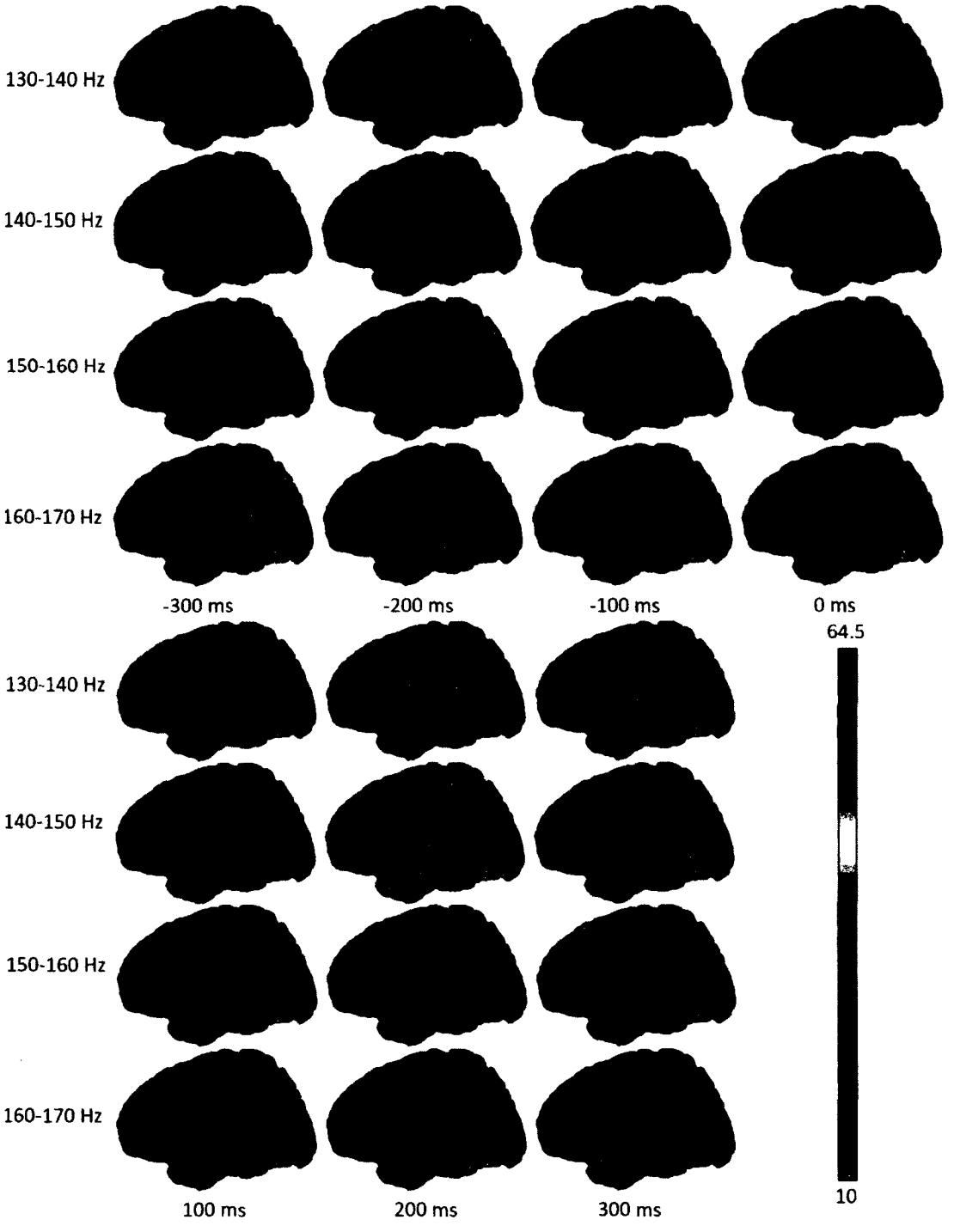
6.1.2 HIGHER SPECTRAL RESOLUTION: THE GAMMA SUB-BANDS

Since the high gamma band is a very wide band of features, i.e., between 70 and 170 Hz, it may be possible that this large feature space holds discriminative information for different components of speech production and perception in the human cortex. It has been shown in an earlier work that different ECoG gamma sub-bands can discriminate between different speech tasks such as reading, listening and speaking, in different cortical locations [93]. This suggests that the ECoG high gamma band should be investigated as a large non-uniform and heterogeneous band of frequencies, in order to utilize smaller spectral resolutions for characterizing different production-based and perception-based speech representations. Furthermore, certain gamma sub-bands may be able to better predict certain speech representations, which would improve the decoding model by fine-tuning it to the specific speech representation being predicted. Hence, eight gamma sub-bands between 70 and 170 Hz, in bins of 10 Hz, were extracted, with the exception of 110-130 Hz for removing the influence of the 120 Hz power line interference. The spatio-temporal relationships between these eight sub-bands and the speech power were investigated, as shown in Figure 19.

It can be observed from this figure that certain gamma sub-bands show a better time evolution of cortical activations relative to speech progression than others. The gamma sub-bands between 90 Hz and 110 Hz show distinctly higher activations in the auditory cotex at a 100 ms time latency compared to other bands. The first four


70-80 Hz
80-90 Hz
90-100 Hz
100-110 Hz
-300 ms
-200 ms
-100 ms
0 ms
70-80 Hz
80-90 Hz
90-100 Hz
100-110 Hz
100 ms
200 ms
300 ms
64.5
10

(a)



(b)

FIG. 19: Spatiotemporal correlations between the eight ECoG high gamma sub-band powers (sub-bands one through four in (a) and sub-bands five through eight in (b)) and the speech power, across seven time latencies relative to the onset of speech.

gamma sub-bands, i.e., the gamma frequencies before 110 Hz, show higher activations in the auditory cortex, Wernicke's area, and the superior temporal gyrus, following speech onset. This suggests that these lower sub-bands hold more information about speech perception than do the higher sub-bands, with the sub-bands between 90 Hz to 110 Hz being the most informative for speech perception.[1] In order to further investigate this discriminative power of the ECoG high gamma activity, these gamma sub-bands were used to re-develop regression models to predict the speech power and the fundamental frequency as discussed in Section 6.2.1.

6.2 OPTIMIZED DECODING MODELS

Optimized decoding models may be developed using more features as inputs in the models or using more optimized model solution techniques. Both these possibilities are explored here, with the use of the gamma sub-band features as a more optimized feature sub-set for the models, and investigating the use of artificial neural networks (ANNS) as a more optimized model solution technique.

6.2.1 OPTIMIZED FEATURES: THE GAMMA SUB-BANDS

Using the eight gamma sub-band powers as features and using the exact same feature selection and linear regression technique, the speech power and the fundamental frequency were predicted, the results of which are shown in Table 6. It should be noted that the feature space, prior to feature selection, is increased 8-fold by this

[1]This analysis was repeated with the silence periods (between the spoken words and sentences that the subjects were silent during) removed from both ECoG gamma sub-band activities and the speech power. The results from this analysis are shown in the Appendix.

method. The results show a vast improvement in prediction of these two basic speech features from the results obtained using the earlier models. This means that using gamma sub-band features instead of the entire gamma band as one feature, in combination with feature selection, helps the model better predict speech power and F0, making the results more significant and with increasing feasibility for practical use. This indicates that certain gamma sub-bands may contain temporal features more suited to the prediction of certain speech representations and these sub-bands are chosen for inclusion in the model in the feature selection stage, which gives us a more optimized model than using temporal features from the whole gamma band. Similarly, these features could possibly be used to predict the other speech representations to improve the modeling results for those representations.

TABLE 6: Correlation coefficients (testing) of the correlation between the actual and the predicted speech power and fundamental frequency from the preparation-based ECoG gamma sub-band powers (combination of ECoG gamma sub-band activities at -200 ms, -100 ms, 0ms). All the models developed were statistically significant, i.e., $p<0.05$ using the randomization test.

Sub.	A	B	C	D	E	F	G	H
Speech Power	0.36	0.59	0.68	0.44	0.32	0.49	0.44	0.55
Fundamental Frequency	0.50	0.73	0.51	0.35	0.20	0.36	0.20	0.49

To further evaluate the differential contributions of the sub-bands in the prediction models, the number of features chosen for each of the eight sub-bands in the feature selection stages of the subject-specific models were evaluated for predicting the speech power. A similar evaluation can also be performed for the models for predicting the fundamental frequency. It should be noted that the features chosen for

each sub-band correspond to the gamma sub-band powers from three temporal lags: -200 ms, -100 ms, and 0 ms, with 0 ms corresponding to the speech onset. Figure 20 shows this evaluation, summed over all the eight subjects. It is observed from this figure that the gamma sub-bands in the frequency range 90-100 Hz and 100-110 Hz contribute the most to the prediction models. The sub-bands in the range 140-150 Hz contribute the third most to the prediction models, followed by the sub-bands in the range 150-160 Hz. Thus, based on this figure, it can be hypothesized that the gamma frequencies between 90-110 Hz hold the most information for predicting the speech power, followed by the gamma frequencies in the range 140-150 Hz, and 150-160 Hz, respectively. This evaluation can also be performed in a subject-specific manner, which was not performed here because the aim is to find generalized trends in sub-band selection across patients.

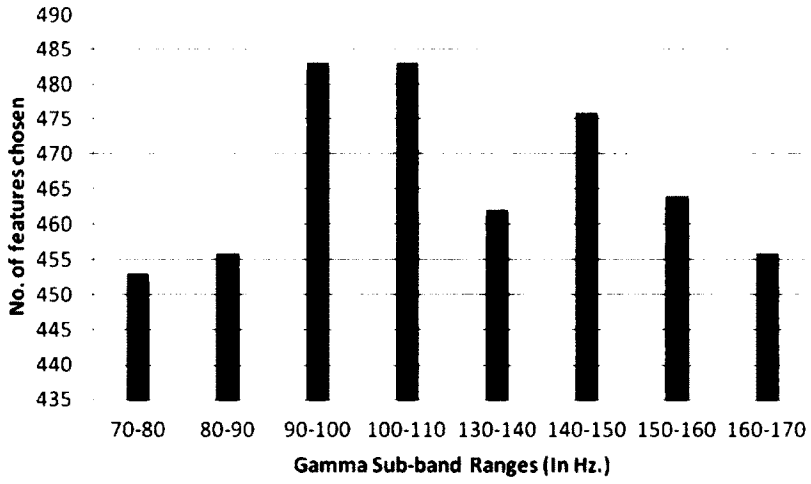


FIG. 20: The number of temporal features chosen for each of the eight gamma sub-bands in the feature selection stage, for modeling the speech power using the gamma sub-band powers, summed over all the eight subjects.

As a more in-depth analysis of the channel selection for the gamma sub-band-based models, the channels selected for each of the eight sub-band for three time lags are shown in Figure 21. The number of features selected in each fold of the cross-validation was restricted to 100, for a more efficient visualization and interpretation. It is observed from this figure, similar to the prior analysis discussed in Chapter 4, that the channels selected were spread out over most of the recording areas, including the primary and secondary auditory areas the showed significant correlations between the different gamma sub-bands and the speech power (see Figure 19).

Based on the above evaluation, new models for predicting the speech power were developed by restricting the gamma sub-band features in the models to the best 2, 3 or 4 sub-bands. The correlations between the actual and the predicted speech power obtained from these models are shown in Figure 22. It is observed from this figure that, in general, the r-values were much lower when using fewer sub-bands as compared to using all the eight sub-bands as inputs to the model. For all the eight subjects, the model with the best performance for predicting the speech power was that developed using all the eight sub-bands as features. Thus, it may be concluded that although the eight gamma sub-bands evaluated have differential contributions in the models for predicting the speech power, all the eight sub-bands combined are important features for the model. The model performance drops drastically when only specific sub-bands, albeit the most important ones, are used as input features in the model.

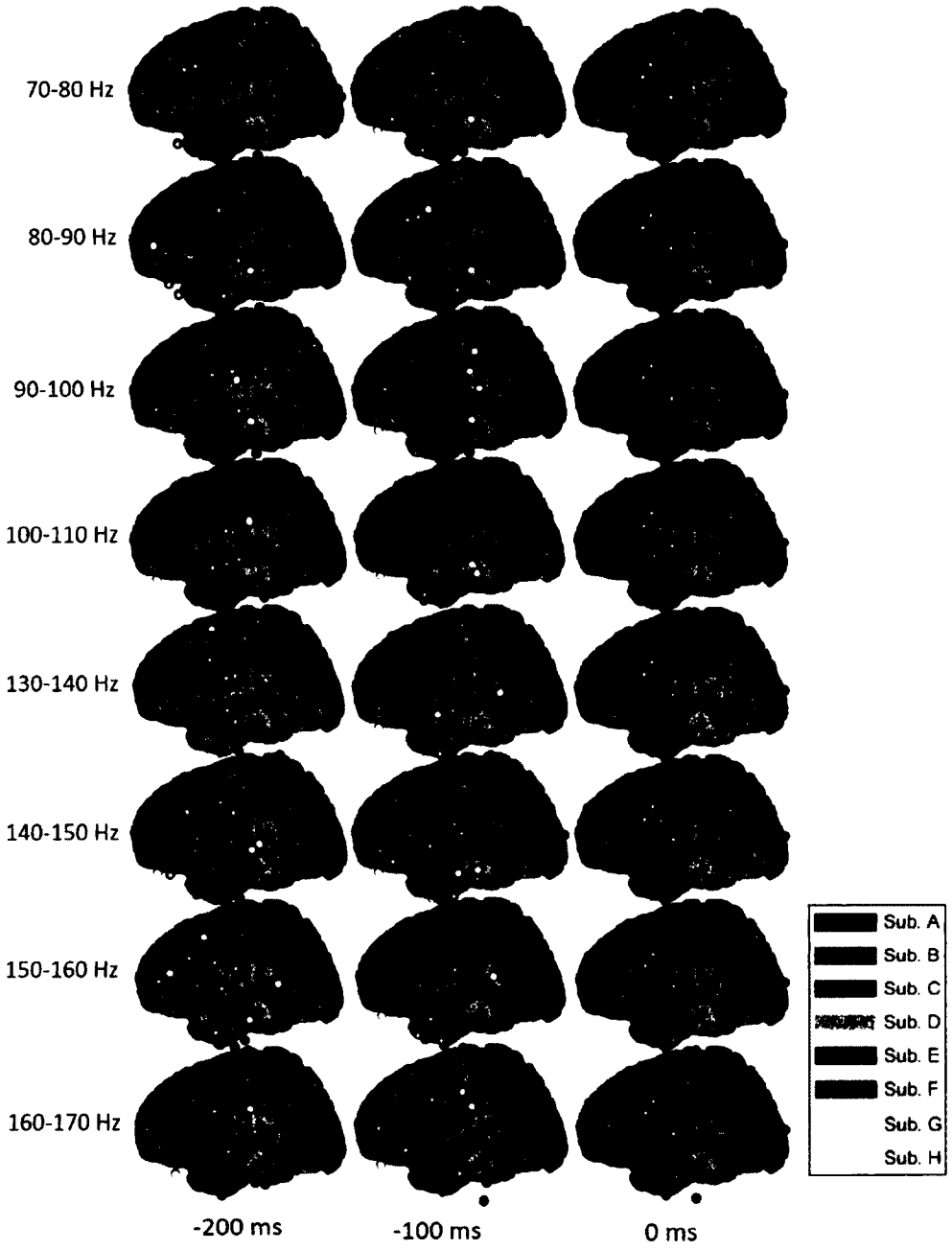


FIG. 21: Channels selected for each of the eight gamma sub-bands, for all the eight subjects, shown on a generic head model, for the preparation-based decoding model, for the three time latencies -200 ms, -100 ms and 0 ms respectively.

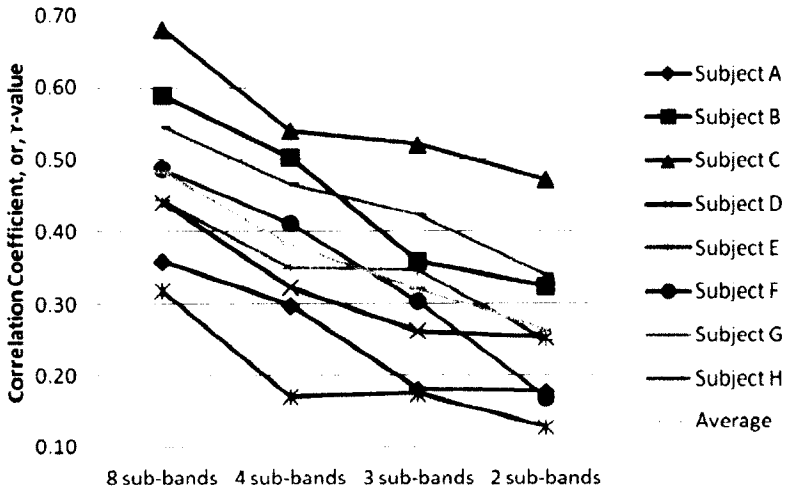


FIG. 22: Correlation between the actual speech power and that predicted by the model, using all the eight gamma sub-bands, the best 4 gamma sub-bands (i.e., 90-100 Hz, 100-110 Hz, 140-150 Hz, 150-160 Hz), the best 3 gamma sub-bands (90-100 Hz, 100-110 Hz, 140-150 Hz) and the best 2 gamma sub-bands (90-100 Hz, 100-110 Hz) respectively.

6.2.2 ARTIFICIAL NEURAL NETWORKS

An artificial neural network (ANN) is a type of machine learning model inspired by biological neural systems, particularly the brain. An ANN is typically an interconnected web of units called neurons, that process and transfer information just as in the human nervous system. A typical ANN consists of an input layer with multiple inputs being fed to the ANN, one or more hidden layers each consisting of one or more neurons, and an output layer that represents one or more outputs of the ANN. Figure 23 shows an example of the simplest form of an ANN, with three inputs, one hidden layer consisting of four neurons and an output layer with two outputs. This is a three-layer feed-forward network, the type of ANN used in the dissertation.

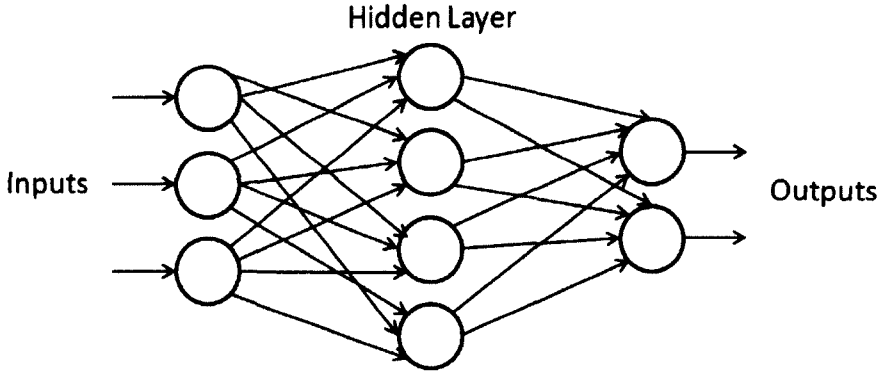


FIG. 23: Basic structure of a simple artificial neural network

This type of ANN tries to map the output (or outputs) based on a combination of weighted inputs and weighted outputs of activation functions acting on the inputs as shown in the equation below.

$$y = W \times \phi(\sum_{i=0}^{D} W_2(i)X(i)) + W_3 \times X \tag{11}$$

Here, W are the weights from the hidden layer to the output layer, W_2 are the weights from the input layer to the hidden layer, and W_3 are the weights from the input layer directly to the output layer. Furthermore, ϕ is the activation function that acts on the variables coming out of the hidden layer, before they are passed on to the output layer. A large number of activation functions are popularly used in ANNs, including the sigmoid function, which has been used in this study. The weights, W, W_2 and W_3 can be determined by either supervised or unsupervised learning methods. Here, supervised learning has been employed, where the output being predicted by the ANN, is used to train the ANN. Most of the supervised learning methods use some form of the gradient descent algorithm to compute the weights, using backpropagation to compute the gradients. This is done by minimizing the cost function with respect to to the network parameters, or the weights, by taking steps in a gradient related direction. In this particular study, the conjugate gradient descent with backpropagation is used for training the ANN, where in the optimum solution is searched for by taking consecutive steps in conjugate directions.

ANNs have been popularly used to decode numerous variables from ECoG activity, ranging from motor and speech imagery to neural spikes [4, 94]. This forms the basis for using the ANN for this particular data set. Furthermore, as compared to linear regression, the ANN incorporates some amount of non-linearity in the models generated and, thus, may be able to better approximate the non-linearities which exist in the data, but are not being adequately captured by the linear regression technique used in Chapters 4 and 5. The ANN is used to predict the speech power and

the fundamental frequency, as described in the following two sections respectively, for the eight subjects using the ECoG gamma activity as the model input.

Prediction of the Speech Power Envelope using Artificial Neural Networks

Eight subject-specific ANNs were developed, for predicting the speech power from ECoG gamma power. The inputs to the ANN were ECoG gamma sub-band powers from -200ms, -100ms and 0 ms (relative to speech onset) across all the channels (50-120), after feature selection. The method used for feature selection was a correlation-based feature selection technique as described in Chapter 3. The performance of the ANN was tested using the Pearson correlation coefficient between predicted and actual speech power. The significance of the correlations were tested using a randomization test to determine the p-value of correlation, identical to that used for the other correlations in this dissertation. The number of hidden layers used in the ANNs was restricted to 1 for reduced complexity. The number of neurons in the hidden layer were varied between 20, 50, 100, 200 and 300 to build five different ANNs and compare the performance of the ANN for this varying number of neurons. A sigmoid transfer function for the hidden layer and a linear transfer function for the output layer were used. The method of conjugate gradient backpropagation was used for optimizing the ANN. 80% of the data was used to train the ANN, 10% to validate it and the remaining 10% to test its performance. Only the testing correlations are reported here. The results were averaged over 10 runs of the ANN to ensure stability.

Figure 24 shows the performance of the ANN for predicting the speech power, versus the number of neurons in the hidden layer. It is observed from this figure

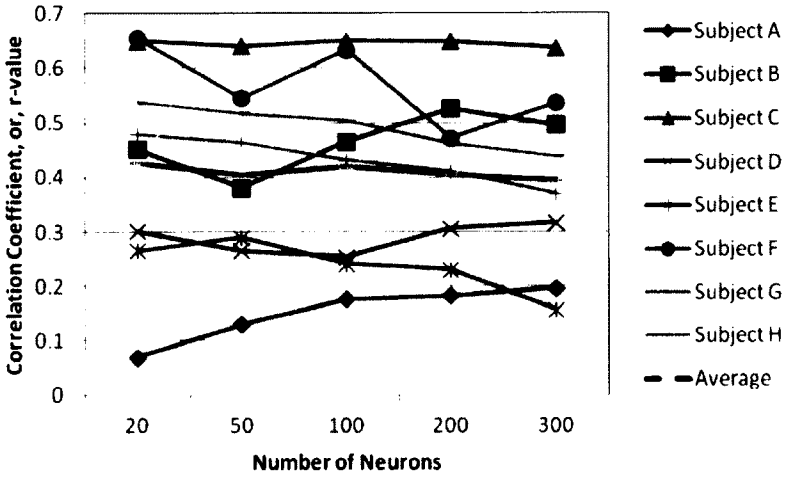


FIG. 24: Correlation between the actual and the speech power predicted by the ANN, versus the number of neurons in the hidden layer of the ANN.

that for some subjects, such as Subjects A and D, the best performance is achieved for the largest number of neurons in the hidden layer, while for some others such as subjects F, G and H, the best performance is achieved for the lowest number of hidden neurons. For subjects C and E, the highest performance is achieved for an intermediate number of neurons, i.e., 100 and 50 respectively. On average, the performance of the ANN does not vary much with the number of hidden neurons. Hence, it may be sufficient to use a network with as little as 20 hidden neurons as the complexity of the network increases significantly with the size of the hidden layers. All the models developed using the ANN were statistically significant, and the r-values obtained for predicting the speech power are on par with the other models examined.

The prediction of the speech power using ANNs is compared to that using the linear regression technique, as shown in Figure 25. Here, the ANN with the optimum

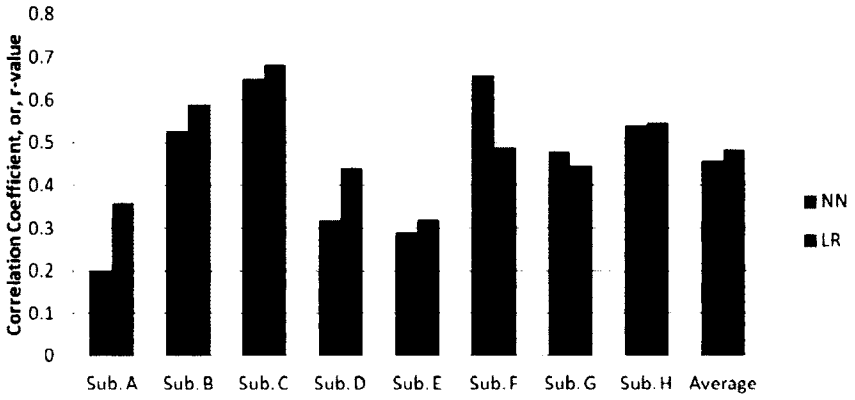


FIG. 25: Comparison of the performance of the best ANN and the linear regression method, for speech power predictions, for all the eight subjects and on an average.

number of hidden neurons is chosen for comparison to linear regression, for each subject. It can be observed from this figure that for most of the subjects (6 out of 8), the linear regression method performs marginally better than ANNs for the prediction of speech power from ECoG gamma power. This may be because the speech power is a relatively simple speech feature, and has been shown to be linearly represented in the human cortex [50]. Thus, linear techniques such as linear regression may be sufficient to encapsulate the complexities that exist in the speech power envelope, and predict it from cortical activity. More complex non-linear techniques may not be necessary for decoding speech envelope representations from cortical activity.

Prediction of the Fundamental Frequency using Artificial Neural Networks

Eight subject-specific ANNs were developed, for predicting the fundamental frequency from ECoG gamma power. The identical feature selection, model selection,

and training procedure was used as described in the previous section.

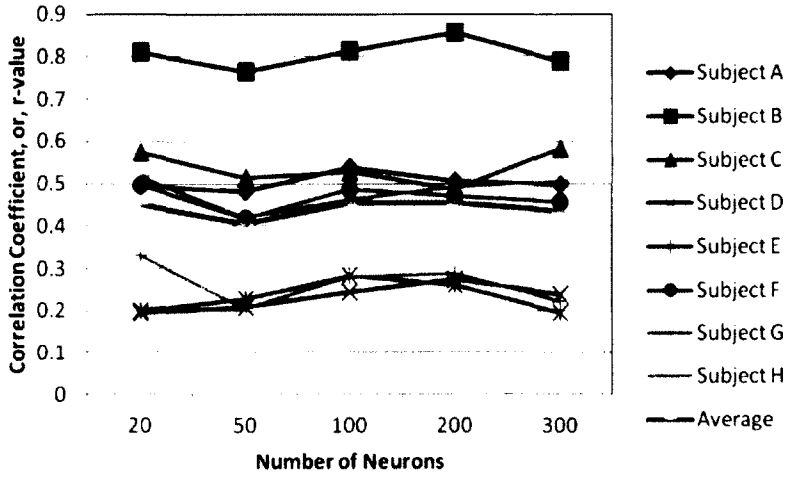


FIG. 26: Correlation between the actual and the fundamental frequency predicted by the ANN, versus the number of neurons in the hidden layer of the ANN.

Figure 26 shows the performance of the ANN for predicting the fundamental frequency, versus the number of neurons in the hidden layer. It may be observed from this figure that, for most of the subjects, a relatively stable ANN performance was observed across the size of the hidden layer. Again, an ANN with as little as 20 hidden neurons may be sufficient for predicting the fundamental frequency from ECoG gamma power. All the models developed using the ANN were statistically significant, and the r-values obtained for predicting the fundamental frequency, are on par with the other models examined.

A comparison of the ANN and linear regression techniques for the prediction of the fundamental frequency is shown in Figure 27. Here, the ANN with the optimum number of hidden neurons, for each subject, is chosen for comparison. It may be

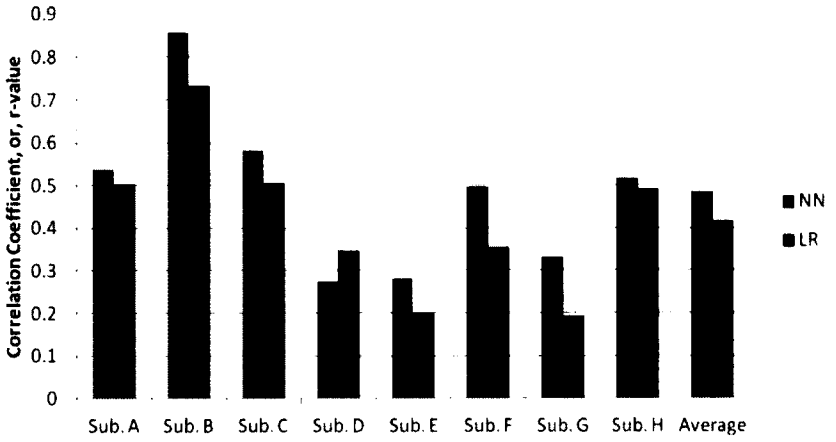


FIG. 27: Comparison of the performance of the best ANN and the linear regression method, for the fundamental frequency predictions, for all the eight subjects and on an average.

observed from this figure that for most of the subjects (7 out of 8), the ANN performs better than linear regression for the prediction of the fundamental frequency from ECoG gamma power. On average, the ANN outperforms the linear regression technique, in the prediction of the fundamental frequency from ECoG gamma activity. This may be because the fundamental frequency of speech, or the pitch, is a slightly more complex speech feature than the speech power envelope. Thus, linear techniques may be insufficient to encapsulate the complexities that exist in the pitch, and predict it from cortical activity. Since the ANNs developed here are more complex non-linear techniques, they better decode pitch information from gamma band activity in the human cortex.

CHAPTER 7

CONCLUSIONS

The focus of this dissertation was on developing methods used to characterize and decode various speech representations commonly used in modern speech recognition systems, directly from cortical activity. This chapter concludes this dissertation with a summary of the main contributions and several possible future directions of this research.

7.1 MAIN CONTRIBUTIONS

Previous studies on ECoG-based speech decoding have attempted to decode vowels, phonemes, words, spectrograms, envelopes and other speech features from the ECoG activity. The primary contribution of this dissertation is that it attempts to decode speech representations commonly used in automatic speech recognition systems, directly from ECoG activity. These speech representations are more complex speech features than those which were previously decoded using ECoG activity, which leads to a more in-depth understanding of speech representation in the human cortex. Thus, this work develops a natural extension of automatic speech recognition techniques to neurological data, and provides a step toward neural speech prostheses, with an automatic speech recognition system as an intermediate step. Because of the recent boom in the development of automatic speech recognition, the results from this dissertation can serve as input to a state-of-the-art speech recognition system in

order to create technologies such as an ECoG-decoded speech synthesis system or an ECoG-based speech-to-text system.

Secondly, most existing ECoG speech and language studies use relatively discontinuous speech production and perception tasks, such as cued word repetition tasks, in order to understand speech representation in the human cortex. This work uses a more fluent and continuous speech production and listening task using familiar stimuli, which is a better representation of natural verbal communication. This research represents a step towards a more practical speech neural prosthesis by allowing the interpretation of neurological activity that is similar to what is expected in normal conversation.

Thirdly, this study utilizes both speech production and speech perception tasks simultaneously, which were studied separately in the majority of earlier studies. This simultaneous study of speech production and perception provides a clearer idea of the areas activated in the cortex during the temporal progression of speech planning, production, articulation, and perception.

Furthermore, due to the recording technique and the experimental design used, it became possible to use a temporal resolution as high as 20 ms. This provides a detailed and precise understanding of the cortical time evolution of speech production and planning. The results of this dissertation provide a very detailed electrophysiological representation of the continuous speech process. Additionally, the results highlight the common neural correlates of overt speech across different linear and non-linear representations of speech, both of which are unique contributions of this

study.

Finally, most existing ECoG speech studies use a wide band of gamma frequencies for studying the cortical involvement in speech tasks as well as for decoding speech components from the cortex. While one earlier study showed that different gamma sub-bands can discriminate between different speech tasks [93], this dissertation expands on this by using eight gamma sub-bands to spatio-temporally characterize the cortical areas involved in speech production and perception, as shown in Figure 19 (Pages 80-81). It is well-demonstrated that unique gamma sub-bands contribute differently to temporal activations during speech production and perception. This led to the development of improved models for decoding the speech power and the fundamental frequency, as shown in Chapter 6. These novel decoding models used the best gamma sub-bands for a particular time latency as the input features for the decoding model, as opposed to the traditional use of the entire wide band of gamma frequencies as model inputs for predicting speech. These novel ECoG features, when combined with either a linear regression or a neural network technique for model solution, led to improved ECoG-based decoding models for predicting the speech power and the fundamental frequency.

7.2 FUTURE DIRECTIONS

Although the research in this dissertation holds significant value as a step towards an ECoG-based speech neuroprosthesis, there are several possible future directions of this research, as discussed in this section.

7.2.1 MODEL IMPROVEMENT

The models that were used to predict the speech power, the fundamental frequency, formants, the LPC coefficients, the MFC coefficients, and the PLP coefficients were very simple models using standard features and feature selection techniques to predict these representations from cortical activity. In practice, more sophisticated models which give better prediction results may be required. There are three main areas of improvement for the models: the features used in the model, the feature selection method used and the modeling technique used.

Features used in the Models

In the models used in Chapters 4 and 5 to predict speech representations, only the ECoG gamma band powers from all the channels were used as features. However, other sets and subsets of features may be more useful for modeling these representations. This is discussed to a certain extent in Chapter 6 where the gamma band is further divided into sub-bands in order to build more optimized models. It was observed from this analysis that the use of the gamma sub-band powers instead of the traditional whole gamma band power as the model inputs led to significant improvement in the decoding models for the prediction of the speech power envelope and the fundamental frequency, the two most important components of the speech signal. Similarly, these sub-band features may also lead to more optimized models for the prediction of the other speech representations discussed in Chapters 4 and 5. Other features may also be used for model improvement, such as various linear

and non-linear transformations of the temporal or spectral ECoG power, which are better correlated with the speech representation being predicted.

Feature Selection Method

In the models developed in Chapters 4 and 5, only correlation-based feature selection was performed, which chose the ECoG features that were best correlated with the speech representation being predicted, i.e., the top 25% of the features. Stepwise selection was also attempted to be used for prediction of the speech power; however, this did not lead to any improvement in the prediction results. More feature selection methods may also be explored, such as mutual information-based feature selection, or the Maximum Relevance Minimum Redundancy (MRMR) technique, to see if the models can be further improved using other feature selection methods. In the mutual information-based feature selection, also known as the Maximum Relevance (MR) technique, the features which have the highest mutual information with the predictor variable, are selected to be included in the model. This is similar to the correlation-based feature selection, although mutual information is a more popular measure used in information theory to quantize the dependency between variables. The MRMR technique is an improvement over the MR technique in the sense that it uses a pre-filtering method to remove redundant features, so that features which are not only of maximum relevance but also of least redundancy are selected. Other techniques such as the Principal and Independent Component Analyses, Subspace Learning etc., can also be investigated as possible dimensionality reduction techniques.

Model Solution Technique

The models developed for predicting the speech representations from ECoG activity were primarily solved using linear regression. Linear regression is a relatively simple method of solving for models by assuming a linear relationship between the model input and output, and minimizing the least squares error. However, in practice, the relationship between ECoG and various speech representations may not be linear and may involve at least a certain degree of non-linearity. The use of artificial neural networks for predicting the speech power and the fundamental frequency was explored in Chapter 6. It was observed from this analysis that the use of artificial neural networks led to an improvement in the decoding models for the prediction of the fundamental frequency or the pitch of the speech signal, but did not provide an improvement for the prediction of the speech power envelope. This may primarily be attributed to the fact that the speech power envelope is a relatively simple linear speech feature, encoded linearly in the human cortex [50]. Hence, a linear regression, which leads to a linear model, is better for predicting this simple speech representation. However, the fundamental frequency of speech is a slightly more complex speech feature, and cannot be expressed as effectively using a simple linear combination of cortical features. Hence, a more non-linear and complex decoding model, as is solved for using the neural network, is better able to decode this speech representation. Similarly, the use of artificial neural networks may also be explored for predicting the other production-based and perception-based speech representations discussed in Chapters 4 and 5. Other modeling techniques such as Kalman filtering

or other types of non-linear regression may also be used to investigate if they improve the prediction of the speech representations. Kalman filtering is a recursive algorithm that uses a series of observations over time to produce estimates of the underlying model relating the input and output data. It is known to be more precise than other modeling techniques because it recursively improves the estimation of the model parameters by using multiple observations over time instead of a single observation. Non-linear regression is a form of regression analysis in which the output variable is assumed to be non-linearly related to the input variables, following a certain non-linear representation. The parameters of this non-linear representation are solved for using successive approximations.

7.2.2 SPEECH RECONSTRUCTION

An application of this research would be to include a method to close the loop, i.e., to use the ECoG-decoded speech representations to reconstruct the acoustic speech signal itself, or to be used as inputs to existing speech recognition systems for the development of an ECoG-based automatic speech recognition system. Existing speech recognition techniques may be applied on the ECoG-decoded speech representations, to lead to various ECoG-based speech recognition applications. The speech features and representations decoded from the ECoG activity may also be used to reconstruct speech, either by inverting the transformations used to obtain these representations, or by combining different decoded speech features and using existing models to recover the speech signal. Based on the different speech representations discussed in the previous chapters, alternative ways to reconstruct speech are discussed here.

Using Production-based Representations

The speech power and the fundamental frequency decoded from the ECoG activity could be used to obtain an amplitude and frequency modulated sinusoidal waveform, which would represent the reconstructed waveform of the speech that was being spoken by the subjects. The waveform would be amplitude-modulated by the decoded speech power values, varying over time, and frequency modulated by the decoded fundamental frequency values, varying over time. Once this basic waveform is obtained, the formants can then be used to create additional sinusoidal waves with frequencies equal to the formant frequencies and these sinusoids can be integrated into the basic speech waveform to add further complexity to the reconstructed waveform, so that it is a better approximation of the actual speech wave. Together, these sinusoids will replicate the estimated frequency and amplitude patterns of the resonance peaks of speech. However, this sine-wave speech may not be very intelligible, as it may be stripped of the acoustic components of speech. The decoded speech power, fundamental frequency and formants may have estimation errors which may lower the intelligibility of the speech as well. These are issues which have to be tackled if using this approach to reconstruct the speech waveform.

Speech may also be reconstructed using the linear prediction coding coefficients that were decoded from ECoG activity, using methods similar to those typically used to obtain speech from the regular linear predictive coding coefficients [95]. The 10 LPC coefficients can be used to predict the current speech sample from the 10 preceding speech samples, using Equation 3. The ECoG-decoded LPC coefficients

can also be used in combination with the ECoG-decoded fundamental frequency, or, the pitch period and the speech power, to get a better estimate of the actual speech signal, as is shown in Figure 12 (Page 57). If this predicted speech sample is combined with an excitation signal, which is an impulse train with period equal to the pitch period (obtained from the decoded fundamental frequency) and amplitude equal to the speech power values for voiced speech and random noise for unvoiced speech, then a better prediction of the current speech sample is achieved. This predicted speech signal should then be low-pass filtered to obtain the continuous speech wave.

Using Perception-based Representations

The mel frequency cepstral coefficient vectors decoded from the ECoG activity can be inverted back to obtain a smoothed estimate of the speech wave, as has been shown in previous work [96]. A theorem known as the Wiener-Khintchine theorem and linear predictive analysis can transform the MFCC vectors into the vocal tract filter coefficients or the LPC coefficients. The same procedure which is used to transform the LPC coefficients into an estimate of the speech signal, as described in the previous section, may then be used to obtain the transformation of these coefficients into the speech signal. The RASTA-PLP toolbox which was used to obtain the MFCC coefficients from the original speech signal, for the purpose of this research, also contains an algorithm for MFCC inversion. This algorithm does a step-by-step inversion procedure in which the steps used to obtain the MFCC coefficients are reversed in an attempt to reconstruct the speech signal from the MFCC vectors. However, since not all the steps used to obtain the MFCC coefficients are reversible,

only an approximation of the speech signal is obtained, and there is an error of estimation. This error will only increase if we use ECoG-decoded MFCCs instead of the original MFCCs for speech reconstruction, as the error in prediction will also be propagated in this inversion process. The RASTA-PLP toolbox also contains a method for inverting the PLP computation to obtain the speech signal from the ECoG decoded PLP coefficient. This is similar to the one used for MFCC inversion, wherein a Bark scale inversion is used instead of a mel scale inversion. Additional steps to make up for the critical band filtering and power compression steps in the PLP computation are also performed. However, for both the MFCC and the PLP inversion techniques, the estimation errors and the problem of the propagation of the prediction errors will have to be resolved if speech reconstruction using these methods is pursued.

7.2.3 COVERT/IMAGINED SPEECH

As was mentioned in Section 3.2, during the time of data collection, ECoG data was also recorded during a covert speech task, in which the subjects imagined saying the same displayed text as was used for the overt speech task. The models developed here were developed for the overt speech task, using the overt speech data. Optimized versions of these models may also be applied to the ECoG data recorded during the covert task. This may be done by following the same pre-processing and filtering for the covert task ECoG signals as was used for the overt task ECoG signals. Some previous studies have attempted to predict components of imagined speech directly from ECoG activity [35, 44, 62]. However, these studies have used relatively simple

speech components for decoding. The speech representation-based decoding models developed here would lead to novel models for decoding covert speech from ECoG activity. The ability to decode continuous segments of imagined or covert speech with a significant level of accuracy will be a monumental addition to the field of ECoG-based speech BCIs and is the ultimate goal of this research.

7.2.4 OTHER FUTURE DIRECTIONS

Some of the existing ECoG studies use micro-ECoG grids, i.e., ECoG grids with more densely spaced electrodes than regular ECoG grids, in order to decode speech [32–35]. These micro-ECoG grids offer higher spatial resolutions than regular ECoG grids and hence, can be beneficial for localization and decoding of speech components more effectively than standard clinical ECoG grid spacing. Other recent speech studies have analyzed intra-cortical electrode recordings that penetrate the cortex in order to decode various components of speech such as imagined vowels and phonemes, directly from cortical activity [97, 98]. However, human research using micro-ECoG and intracortical electrodes has been limited due to the lack of clinical applicability of these electrodes. Nevertheless, it can be argued that the addition of micro-ECoG to standard clinical grids poses insignificant risk to the patient, and that it may even benefit seizure localization and cortical mapping. For these reasons, it is expected that micro-ECoG will continue to gain clinical acceptance, which will in turn benefit future language decoding and general neuroscience studies [11].

The ability to modulate pitch, intensity, speaking rate, syllabic stress, and rhythm

is an important part of communication. These variables can combine to convey emotional affect and linguistic information beyond standard speech production. Ideally, a practical neuroprosthesis for communication would decode the combination of speech output, intended emotion, motor plans for facial expressions, and intended patterns of intonation and stress together from acquired neural signals for natural speech output. In addition, a practical prosthesis should be able to distinguish between intended speech output and inner speech (e.g., differences between overt and covert speech) to allow potential users a way to monitor and filter their speech output. This description of a neural prosthesis for communication is far beyond the capabilities of any current speech decoding efforts. In addition, some intended users of speech and language neuroprosthetics may possess varying cognitive and motor abilities making it difficult for them to control such a complex neuroprosthetic device. However, a natural communication prosthesis with a direct link to the brain should possess these abilities to offer users the most complete communication experience possible [11].

Finally, while studies that analyze ECoG signal recordings provide invaluable information toward understanding the complex processes involved in speech production and auditory processing, the present study and other existing studies have examined recordings obtained over a short duration in a medical setting. Typically, recordings are obtained over the course of a few minutes to several days while the subjects are stationary, confined to a hospital bed, and interacting with one speaker. In contrast, natural speech communication occurs in the presence of many additional

factors, including visual scene changes and additional motor movements during activities such as walking, nonverbal communication including gestures and eye gaze maintenance, interaction with multiple and varying communication partners, and conversation maintenance and repair after interruption or misunderstanding during lengthy communication exchanges. In order to develop a neural prosthesis for speech and language that successfully functions in natural settings, these factors must be examined in future studies. Chronically implanted electrodes would allow for ECoG recordings in natural communication environments and represent one way to investigate these factors in communication. Few studies have been conducted on the safety and efficacy of long-term electrode implantation in humans, or the long-term placement of ECoG grids. However, early evidence from recent studies of the stability of long-term impedance in chronic subdural electrodes concluded that impedance was stable up to one year after implantation [37, 38]. In addition, these studies reported few adverse effects resulting from chronically implanted electrodes. This highlights the feasibility of the use of chronically placed subdural electrodes for a more natural and practical speech-based neuroprosthesis [11].

7.3 DISCUSSION

The results obtained in this research will pave for the way for future ECoG-based neural speech prosthesis systems, which will be immensely valuable in the neural engineering and communications disorders fields, as well as the healthcare industry. It will greatly benefit a large section of patients who are paralyzed, locked-in or semi locked-in, and have lost all means of communicating with the outside

world. The novel spatio-temporal characterization results discussed in this dissertation provide detailed insights into how speech processing progresses in the human brain, starting from planning to production to perception of self-vocalizations. Although the decoding accuracies with which most of the speech representations are predicted from ECoG activity are not yet practical, the optimized models for predicting the speech power envelope and the fundamental frequency (pitch information) show great promise. These results represent a contribution not only to existing research in the field of brain-computer interfaces, but also to the expanding field of research in speech-based neural decoding. The results from this dissertation may also be extended to decoding covert speech from cortical activity, which is the ultimate objective in the development of a practical ECoG-based speech neuroprosthesis.

BIBLIOGRAPHY

[1] R. C. Ficke, "Digest of data on persons with disabilities." 1992.

[2] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control." *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, Jun 2002.

[3] J. R. Wolpaw and E. W. Wolpaw, "Brain-computer interfaces: something new under the sun," *Brain-computer Interfaces: Principles and Practice*, pp. 3–12, 2012.

[4] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain-computer interface using electrocorticographic signals in humans." *Journal of Neural Engineering*, vol. 1, no. 2, pp. 63–71, Jun 2004.

[5] P. Brunner, A. L. Ritaccio, J. F. Emrich, H. Bischof, and G. Schalk, "Rapid communication with a p300 matrix speller using electrocorticographic signals (ecog)," *Frontiers in Neuroscience*, vol. 5, p. 5, 2011.

[6] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz, "Cortical control of a prosthetic arm for self-feeding," *Nature*, vol. 453, no. 7198, pp. 1098–1101, 2008.

[7] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal

ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, 2006.

[8] G. Bin, X. Gao, Y. Wang, Y. Li, B. Hong, and S. Gao, "A high-speed BCI based on code modulation VEP," *Journal of Neural Engineering*, vol. 8, no. 2, p. 025015, 2011.

[9] C. M. Reed and N. I. Durlach, "Note on information transfer rates in human communication," *Presence: Teleoperators and Virtual Environments*, vol. 7, no. 5, pp. 509–518, 1998.

[10] X. Pei, J. Hill, and G. Schalk, "Silent communication: toward using brain signals." *IEEE Pulse*, vol. 3, no. 1, pp. 43–46, Jan 2012.

[11] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, "Progress in speech decoding from the electrocorticogram," *Biomedical Engineering Letters*, vol. 5, no. 1, pp. 10–21, 2015.

[12] B. N. Pasley and R. T. Knight, "Decoding speech for understanding and treating aphasia." *Progress in Brain Research*, vol. 207, pp. 435–456, 2013.

[13] C. J. Price, "A review and synthesis of the first 20 years of pet and fmri studies of heard speech, spoken language and reading." *NeuroImage*, vol. 62, no. 2, pp. 816–847, Aug 2012.

[14] G. A. Ojemann, "Cortical organization of language." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, vol. 11, no. 8, pp. 2281–2287, Aug 1991.

[15] P. Broca, "Perte de la parole, ramollissement chronique et destruction partielle du lobe antérieur gauche du cerveau," *Bulletin de la Société Anthropologique*, vol. 2, pp. 235–238, 1861.

[16] C. Wernicke, *Der aphasische symptomenkomplex.* Springer, 1974.

[17] G. Hickok and D. Poeppel, "The cortical organization of speech processing." *Nature Reviews. Neuroscience*, vol. 8, no. 5, pp. 393–402, May 2007.

[18] G. Hickok, "Computational neuroanatomy of speech production." *Nature Reviews. Neuroscience*, vol. 13, no. 2, pp. 135–145, Feb 2012.

[19] E. C. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenberg, D. Barbour, and G. Schalk, "Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task." *Frontiers in Human Neuroscience*, vol. 6, p. 99, May 2012.

[20] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production." *Brain and Language*, vol. 96, no. 3, pp. 280–301, Mar 2006.

[21] C. J. Price, R. J. Wise, E. A. Warburton, C. J. Moore, D. Howard, K. Patterson, R. S. Frackowiak, and K. J. Friston, "Hearing and saying. the functional neuroanatomy of auditory word processing." *Brain : A Journal of Neurology*, vol. 119 (Pt 3), pp. 919–931, Jun 1996.

[22] C. J. Price, "The anatomy of language: contributions from functional neuroimaging." *Journal of Anatomy*, vol. 197 (Pt 3), pp. 335–359, Oct 2000.

[23] J. A. Fiez and S. E. Petersen, "Neuroimaging studies of word reading." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 3, pp. 914–921, Feb 1998.

[24] J. R. Binder, J. A. Frost, T. A. Hammeke, P. S. Bellgowan, J. A. Springer, J. N. Kaufman, and E. T. Possing, "Human temporal lobe activation by speech and nonspeech sounds." *Cerebral Cortex*, vol. 10, no. 5, pp. 512–528, May 2000.

[25] T. M. Talavage, J. Gonzalez-Castillo, and S. K. Scott, "Auditory neuroimaging with fMRI and PET." *Hearing Research*, vol. 307, pp. 4–15, Jan 2014.

[26] L. Y. Ganushchak, I. K. Christoffels, and N. O. Schiller, "The use of electroencephalography in language production research: a review." *Frontiers in Psychology*, vol. 2, p. 208, Sep 2011.

[27] L. D. Sanders and H. J. Neville, "An ERP study of continuous speech processing: I. segmentation, semantics, and syntax in native speakers." *Brain Research. Cognitive Brain Research*, vol. 15, no. 3, pp. 228–240, Feb 2003.

[28] P. Hagoort and C. M. Brown, "ERP effects of listening to speech: semantic erp effects." *Neuropsychologia*, vol. 38, no. 11, pp. 1518–1530, 2000.

[29] P. Indefrey and W. J. M. Levelt, "The spatial and temporal signatures of word production components." *Cognition*, vol. 92, no. 1-2, pp. 101–144, May/Jun 2004.

[30] A. Palmini, "The concept of the epileptogenic zone: a modern look at penfield and jasper's views on the role of interictal spikes." *Epileptic Disorders : International Epilepsy Journal with Videotape*, vol. 8 Suppl 2, pp. S10–S15, Aug 2006.

[31] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals." *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140–154, 2011.

[32] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger, "Decoding spoken words using local field potentials recorded from the cortical surface." *Journal of Neural Engineering*, vol. 7, no. 5, p. 056007, Oct 2010.

[33] T. Blakely, K. J. Miller, R. P. N. Rao, M. D. Holmes, and J. G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids." *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2008, pp. 4964–4967, 2008.

[34] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, "Categorical speech representation in human superior temporal gyrus." *Nature Neuroscience*, vol. 13, no. 11, pp. 1428–1432, Nov 2010.

[35] E. C. Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, and G. Schalk, "Using the electrocorticographic speech network to control a brain-computer interface in humans." *Journal of Neural Engineering*, vol. 8, no. 3, p. 036004, Jun 2011.

[36] A. B. Schwartz, X. T. Cui, D. J. Weber, and D. W. Moran, "Brain-controlled interfaces: movement restoration with neural prosthetics." *Neuron*, vol. 52, no. 1, pp. 205–220, Oct 2006.

[37] K. A. Sillay, P. Rutecki, K. Cicora, G. Worrell, J. Drazkowski, J. J. Shih, A. D. Sharan, M. J. Morrell, J. Williams, and B. Wingeier, "Long-term measurement of impedance in chronically implanted depth and subdural electrodes during responsive neurostimulation in humans." *Brain Stimulation*, vol. 6, no. 5, pp. 718–726, Sep 2013.

[38] C. Wu, J. J. Evans, C. Skidmore, M. R. Sperling, and A. D. Sharan, "Impedance variations over time for a closed-loop neurostimulation device: early experience with chronically implanted electrodes." *Neuromodulation : Journal of the International Neuromodulation Society*, vol. 16, no. 1, pp. 46–50; discussion 50, Jan/Feb 2013.

[39] N. E. Crone, A. Sinai, and A. Korzeniewska, "High-frequency gamma oscillations and human brain mapping with electrocorticography," *Progress in Brain Research*, vol. 159, pp. 275–295, 2006.

[40] T. Yanagisawa, M. Hirata, Y. Saitoh, T. Goto, H. Kishima, R. Fukuma, H. Yokoi, Y. Kamitani, and T. Yoshimine, "Real-time control of a prosthetic hand using human electrocorticography signals." *Journal of Neurosurgery*, vol. 114, no. 6, pp. 1715–1722, Jun 2011.

[41] G. Schalk, K. J. Miller, N. R. Anderson, J. A. Wilson, M. D. Smyth, J. G. Ojemann, D. W. Moran, J. R. Wolpaw, and E. C. Leuthardt, "Two-dimensional movement control using electrocorticographic signals in humans." *Journal of Neural Engineering*, vol. 5, no. 1, pp. 75–84, Mar 2008.

[42] T. Hinterberger, G. Widman, T. N. Lal, J. Hill, M. Tangermann, W. Rosenstiel, B. Schölkopf, C. Elger, and N. Birbaumer, "Voluntary brain regulation and communication with electrocorticogram signals." *Epilepsy & Behavior : E&B*, vol. 13, no. 2, pp. 300–306, Aug 2008.

[43] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception. brazier award-winning article, 2001." *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, vol. 112, no. 4, pp. 565–582, Apr 2001.

[44] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, and G. Schalk, "Spatiotemporal dynamics of electrocorticographic high gamma activity during

overt and covert word repetition." *NeuroImage*, vol. 54, no. 4, pp. 2960–2972, Feb 2011.

[45] N. E. Crone, L. Hao, J. Hart, D. Boatman, R. P. Lesser, R. Irizarry, and B. Gordon, "Electrocorticographic gamma activity during word production in spoken and sign language." *Neurology*, vol. 57, no. 11, pp. 2045–2053, Dec 2001.

[46] A. Sinai, C. W. Bowers, C. M. Crainiceanu, D. Boatman, B. Gordon, R. P. Lesser, F. A. Lenz, and N. E. Crone, "Electrocorticographic high gamma activity versus electrical cortical stimulation mapping of naming." *Brain : A Journal of Neurology*, vol. 128, no. Pt 7, pp. 1556–1570, Jul 2005.

[47] E. Edwards, M. Soltani, L. Y. Deouell, M. S. Berger, and R. T. Knight, "High gamma activity in response to deviant auditory stimuli recorded directly from human cortex." *Journal of Neurophysiology*, vol. 94, no. 6, pp. 4269–4280, Dec 2005.

[48] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation." *Nature*, vol. 495, no. 7441, pp. 327–332, Mar 2013.

[49] K. J. Miller, T. J. Abel, A. O. Hebb, and J. G. Ojemann, "Rapid online language mapping with electrocorticography." *Journal of Neurosurgery. Pediatrics*, vol. 7, no. 5, pp. 482–490, May 2011.

[50] J. Kubanek, P. Brunner, A. Gunduz, D. Poeppel, and G. Schalk, "The tracking of speech envelope in the human cortex." *PloS One*, vol. 8, no. 1, p. e53398, Jan 2013.

[51] E. Edwards, M. Soltani, W. Kim, S. S. Dalal, S. S. Nagarajan, M. S. Berger, and R. T. Knight, "Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex." *Journal of Neurophysiology*, vol. 102, no. 1, pp. 377–386, Jul 2009.

[52] K. V. Nourski, R. A. Reale, H. Oya, H. Kawasaki, C. K. Kovach, H. Chen, M. A. Howard, and J. F. Brugge, "Temporal envelope of time-compressed speech represented in the human auditory cortex." *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, vol. 29, no. 49, pp. 15 564–15 574, Dec 2009.

[53] R. T. Canolty, M. Soltani, S. S. Dalal, E. Edwards, N. F. Dronkers, S. S. Nagarajan, H. E. Kirsch, N. M. Barbaro, and R. T. Knight, "Spatiotemporal dynamics of word processing in the human brain." *Frontiers in Neuroscience*, vol. 1, no. 1, pp. 185–196, Nov 2007.

[54] E. F. Chang, C. A. Niziolek, R. T. Knight, S. S. Nagarajan, and J. F. Houde, "Human cortical sensorimotor network underlying feedback control of vocal pitch." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 7, pp. 2653–2658, Feb 2013.

[55] V. L. Towle, H.-A. Yoon, M. Castelle, J. C. Edgar, N. M. Biassou, D. M. Frim, J.-P. Spire, and M. H. Kohrman, "ECoG gamma activity during a language task: differentiating expressive and receptive speech areas." *Brain : A Journal of Neurology*, vol. 131, no. Pt 8, pp. 2013–2027, Aug 2008.

[56] J. D. W. Greenlee, A. W. Jackson, F. Chen, C. R. Larson, H. Oya, H. Kawasaki, H. Chen, and M. A. Howard, "Human auditory cortical activation during self-vocalization." *PloS One*, vol. 6, no. 3, p. e14744, Mar 2011.

[57] W. Wang, A. D. Degenhart, G. P. Sudre, D. A. Pomerleau, and E. C. Tyler-Kabara, "Decoding semantic information from human electrocorticographic (ecog) signals." *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2011, pp. 6294–6298, 2011.

[58] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans." *Journal of Neural Engineering*, vol. 8, no. 4, p. 046028, Aug 2011.

[59] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone, "Joint spatial-spectral feature space clustering for speech activity detection from ecog signals." *IEEE Transactions on Bio-Medical Engineering*, vol. 61, no. 4, pp. 1241–1250, Apr 2014.

[60] D. Zhang, E. Gong, W. Wu, J. Lin, W. Zhou, and B. Hong, "Spoken sentences

decoding based on intracranial high gamma response using dynamic time warping." *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2012, pp. 3292–3295, 2012.

[61] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all american english phonemes using signals from functional speech motor cortex." *Journal of Neural Engineering*, vol. 11, no. 3, p. 035015, Jun 2014.

[62] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex." *Frontiers in Neuroengineering*, vol. 7, p. 14, May 2014.

[63] M. Zavaglia, R. T. Canolty, T. M. Schofield, A. P. Leff, M. Ursino, R. T. Knight, and W. D. Penny, "A dynamical pattern recognition model of γ activity in auditory cortex." *Neural Networks: The Official Journal of the International Neural Network Society*, vol. 28, pp. 1–14, Apr 2012.

[64] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex." *PLoS Biology*, vol. 10, no. 1, p. e1001251, Jan 2012.

[65] R. Behroozmand and C. R. Larson, "Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback." *BMC Neuroscience*, vol. 12, p. 54, Jun 2011.

[66] A. Parbery-Clark, D. L. Strait, S. Anderson, E. Hittner, and N. Kraus, "Musical experience and the aging auditory system: implications for cognitive abilities and hearing speech in noise." *PloS One*, vol. 6, no. 5, p. e18082, May 2011.

[67] A. Heinrich, B. A. Schneider, and F. I. M. Craik, "Investigating the influence of continuous babble on auditory short-term memory performance." *Quarterly Journal of Experimental Psychology*, vol. 61, no. 5, pp. 735–751, May 2008.

[68] M. Pichora-Fuller, C. Palmer, and R. Seewald, "Audition and cognition: What audiologists need to know about listening," *Hearing Care for Adults*, pp. 71–85, 2007.

[69] L. Deng and D. O'Shaughnessy, *Speech processing: a dynamic and optimization-oriented approach.* CRC Press, 2003.

[70] L. Xu and B. E. Pfingst, "Relative importance of temporal envelope and fine structure in lexical-tone perception (l)," *The Journal of the Acoustical Society of America*, vol. 114, no. 6, pp. 3024–3027, 2003.

[71] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[72] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[73] Y.-Y. Kong, G. S. Stickney, and F.-G. Zeng, "Speech and melody recognition in binaurally combined acoustic and electric hearing," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1351–1361, 2005.

[74] G. Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations.* Walter de Gruyter, 1971, vol. 2.

[75] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[76] A. V. McCree and T. P. Barnwell III, "A mixed excitation lpc vocoder model for low bit rate speech coding," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 4, pp. 242–250, 1995.

[77] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, "Hmm-based audio keyword generation," in *Advances in Multimedia Information Processing-PCM 2004.* Springer, 2005, pp. 566–574.

[78] A. Vasilijević and D. Petrinović, "Perceptual significance of cepstral distortion measures in digital speech processing," *AUTOMATIKA: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, vol. 52, no. 2, pp. 132–146, 2011.

[79] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[80] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.

[81] E. Tsiang, "A cochlea filter bank for speech analysis," in *Proc. International Conference on Signal Processing Applications and Technology*, 1997, pp. 1674–1678.

[82] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7. IEEE, 1982, pp. 1282–1285.

[83] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 920–930, 2006.

[84] P. Talairach and J. Tournoux, "A stereotactic coplanar atlas of the human brain," 1988.

[85] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1034–1043, 2004.

[86] S. L. Hahn, *Hilbert transforms in signal processing.* Artech House, 1996.

[87] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.0. 05)[computer program]. retrieved january 19, 2008," 1992.

[88] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154–161, 2006.

[89] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 2, pp. 221–226, 1968.

[90] D. O'shaughnessy, *Speech communication: human and machine*. Universities press, 1987.

[91] K. Brodmann, "Vergleichende lokalisationslehre der groshirnrinde," *Leipzig: Barth*, 1909.

[92] J. S. Brumberg, D. J. Krusienski, A. Gunduz, P. Brunner, S. Chakrabarti, A. L. Ritaccio, and G. Schalk, "Spatio-temporal evolution of cortical processes related to continuous overt and covert speech production in a reading task," *(in progress)*.

[93] C. M. Gaona, M. Sharma, Z. V. Freudenburg, J. D. Breshears, D. T. Bundy, J. Roland, D. L. Barbour, G. Schalk, and E. C. Leuthardt, "Nonuniform high-gamma (60–500 hz) power changes dissociate cognitive task and anatomy in human cortex," *The Journal of Neuroscience*, vol. 31, no. 6, pp. 2091–2100, 2011.

[94] G. Hellmann, "Multifold features determine linear equation for automatic spike detection applying neural nin interictal ecog," *Clinical Neurophysiology*, vol. 110, no. 5, pp. 887–894, 1999.

[95] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[96] B. Milner and X. Shao, "Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model." in *INTERSPEECH*. Citeseer, 2002.

[97] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, "A wireless brain-machine interface for real-time speech synthesis." *PloS One*, vol. 4, no. 12, p. e8218, Dec 2009.

[98] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy, "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex." *Frontiers in Neuroscience*, vol. 5, p. 65, May 2011.

APPENDIX

This section contains the spatio-temporal plots shown in the dissertation, redone with the silence periods removed from both the ECoG activity and the various speech representations. This was done for a truer representation of speech progression in the human cortex. This dissertation used natural speech spoken by the subjects, which included periods of inactivity (or silence) between words and sentences that the subjects spoke aloud. The silence periods are much lower in amplitude than the speech periods are, and including the silence periods in the spatio-temporal characterization may bias the analysis, and not give us detailed information about the neural bases underlying speech production and perception occuring in the non-silence periods, which is of interest in this study.

First, we investigate spatio-temporal correlation between the ECoG high gamma power and the speech power, only during the periods of speech activity. Figure 28 shows these spatio-temporal correlations between ECoG high gamma power and speech power, shown for time latencies between -300 ms to 300 ms, in steps of 100 ms. We see that the activations are much reduced, on comparison to those obtained in Figure 8 (Page 48). The negative lags show almost no significant activations. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, starting at 0 ms, strengthening at 100 ms, and reducing severely at and after 200 ms.

Second, we study the spatio-temporal correlation between the ECoG high

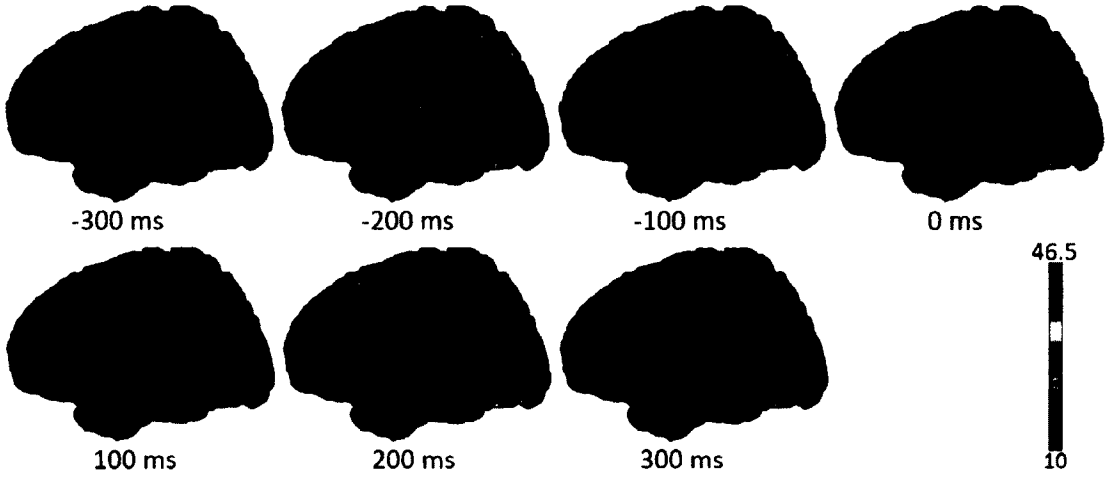


FIG. 28: Spatiotemporal correlations between the speech power and the ECoG high gamma band power, across seven time latencies relative to the onset of speech, with silence periods removed from the analysis.

gamma power and the fundamental frequency, only during the periods of speech activity. Figure 29 shows these spatio-temporal correlations between ECoG high gamma power and speech power, shown for time latencies between -300 ms to 300 ms, in steps of 100 ms. It may be observed that the activations are much reduced, on comparison to those obtained in Figure 11 (Page 55). The negative lags show very less significant activations. However, as compared to those obtained for the speech envelope in Figure 28, some activations are seen in the pre-motor and the motor areas, albeit much reduced. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, starting at 0 ms, strengthening at 100 ms, and disappearing at and after 200 ms.

Next, the spatio-temporal correlation between the ECoG high gamma power and the LPC coefficients is investigated, only during the periods of speech activity. Figure 30 shows these spatio-temporal correlations between ECoG high gamma power

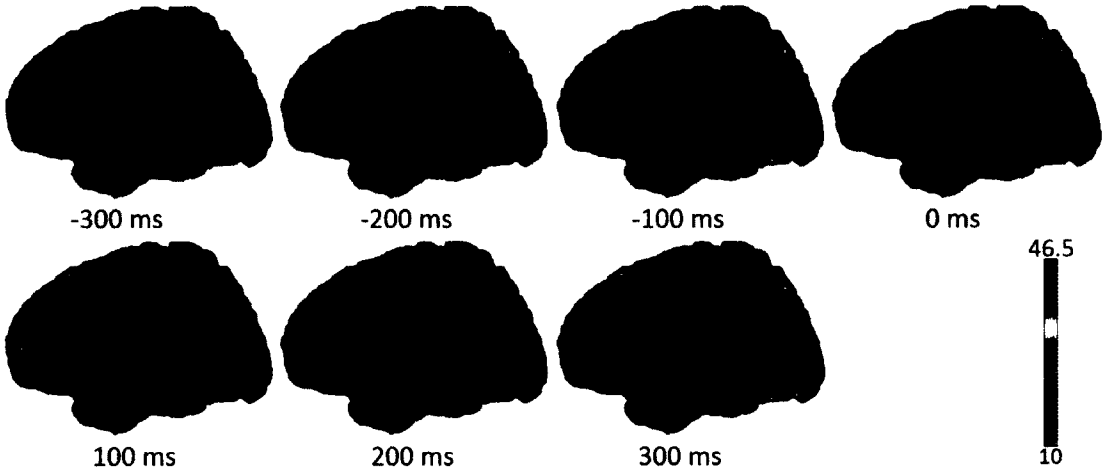


FIG. 29: Spatiotemporal correlations between ECoG high gamma power and the fundamental frequency, across seven time latencies relative to the onset of speech, with silence periods removed from the analysis.

and LPC coefficients for which significant activations were achieved, shown for time latencies between -300 ms to 300 ms, in steps of 100 ms. It is seen that the activations are much reduced, on comparison to those obtained in Figure 13 (Page 60). Also, with the silence periods included in the analysis, the even coefficients (2, 4, 6, 8, 10) were found to be significant. However, with the silence periods removed from the analysis, the first few coefficients (1-4, 6) are found to be significant, which is closer to what we would expect, as the LPC coefficients are based on auto-regressive model with the coefficients arranged in decreasing order of importance. The negative lags show almost no significant activations. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, starting at 0 ms, strengthening at 100 ms, and disappearing at and after 200 ms. The activation of the superior temporal gyrus is found to be the strongest for the second LPC coefficient, which was also found to the most significant in the analysis with silence periods included

(see Figure 13, Page 60).

Further, we study the spatio-temporal correlation between the ECoG high gamma power and the MFC coefficients, only during the periods of speech activity. Figure 31 shows these spatio-temporal correlations between ECoG high gamma power and MFC coefficients for which significant activations were achieved, shown for time latencies between -300 ms to 300 ms, in steps of 100 ms. It is observed that the activations are much reduced, on comparison to those obtained in Figure 14 (Page 66). Also, the first three coefficients were found to be significant in both the analyses (with and without silence periods included). The negative lags show very less significant activations. However, as compared to those obtained for the speech envelope in Figure 28 and most of the LPC coefficients in Figure 30, some activations are seen in the pre-motor and the motor areas, albeit much reduced. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, starting at 0 ms, strengthening at 100 ms, and disappearing at and after 200 ms. These activations are the strongest for the first MFC coefficient, which is the most important, as it corresponds to the Mel filter most important for pitch perception.

Furthermore, the spatio-temporal correlation between the ECoG high gamma power and the PLP coefficients is investigated, only during the periods of speech activity. Figure 32 shows these spatio-temporal correlations between ECoG high gamma power and PLP coefficients for which significant activations were achieved, shown for time latencies between -300 ms to 300 ms, in steps of 100 ms. The activations appear to be much reduced, on comparison to those obtained in Figure

15 (Page 71). Also, only the first coefficient was found to be significant in both the analyses (with and without silence periods included). The negative lags show some very reduced activations in the pre-motor and the motor areas. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, starting at 0 ms, strengthening at 100 ms, and disappearing at and after 200 ms. These activations are the significant for the first PLP coefficient, which is the most important, as it corresponds to the Bark filter most important for loudness perception.

Finally, the ECoG high gamma sub-band characterization was repeated, taking into account only the periods of speech activity. Figure 33 shows these spatio-temporal correlations between ECoG high gamma sub-bands and the speech power, for time latencies between -300 ms to 300 ms, in steps of 100 ms. The activations are much reduced, on comparison to those obtained in Figure 19 (Page 80-81). A clear distinction between the involvement of the different gamma sub-bands for speech production and perception cannot be made from these low activations. The negative lags show some very low activations in the pre-motor and the motor areas, for only the sub-bands 80-90 Hz and 150-160 Hz. Activation in the auditory areas, i.e., the superior temporal gyrus is found mostly at the positive lags, starting at 0 ms, strengthening at 100 ms, and disappearing at and after 200 ms. These activations are in general very low and do not lead to any conclusive interpretations.

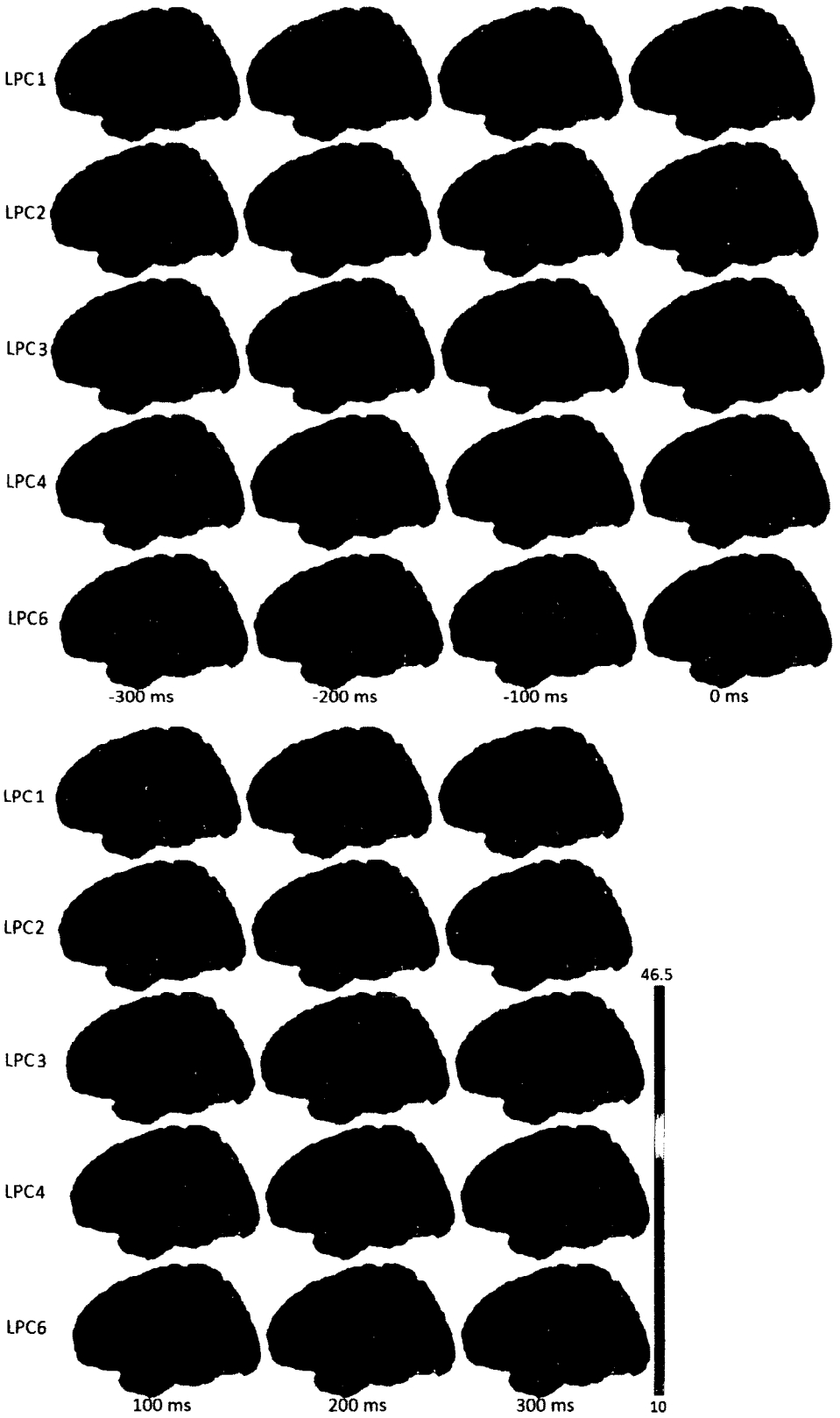


FIG. 30: Spatiotemporal correlations between ECoG high gamma power and the significant LPC coefficients (shown in the five rows respectively) across seven time latencies relative to the onset of speech, with silence periods removed from the analysis.

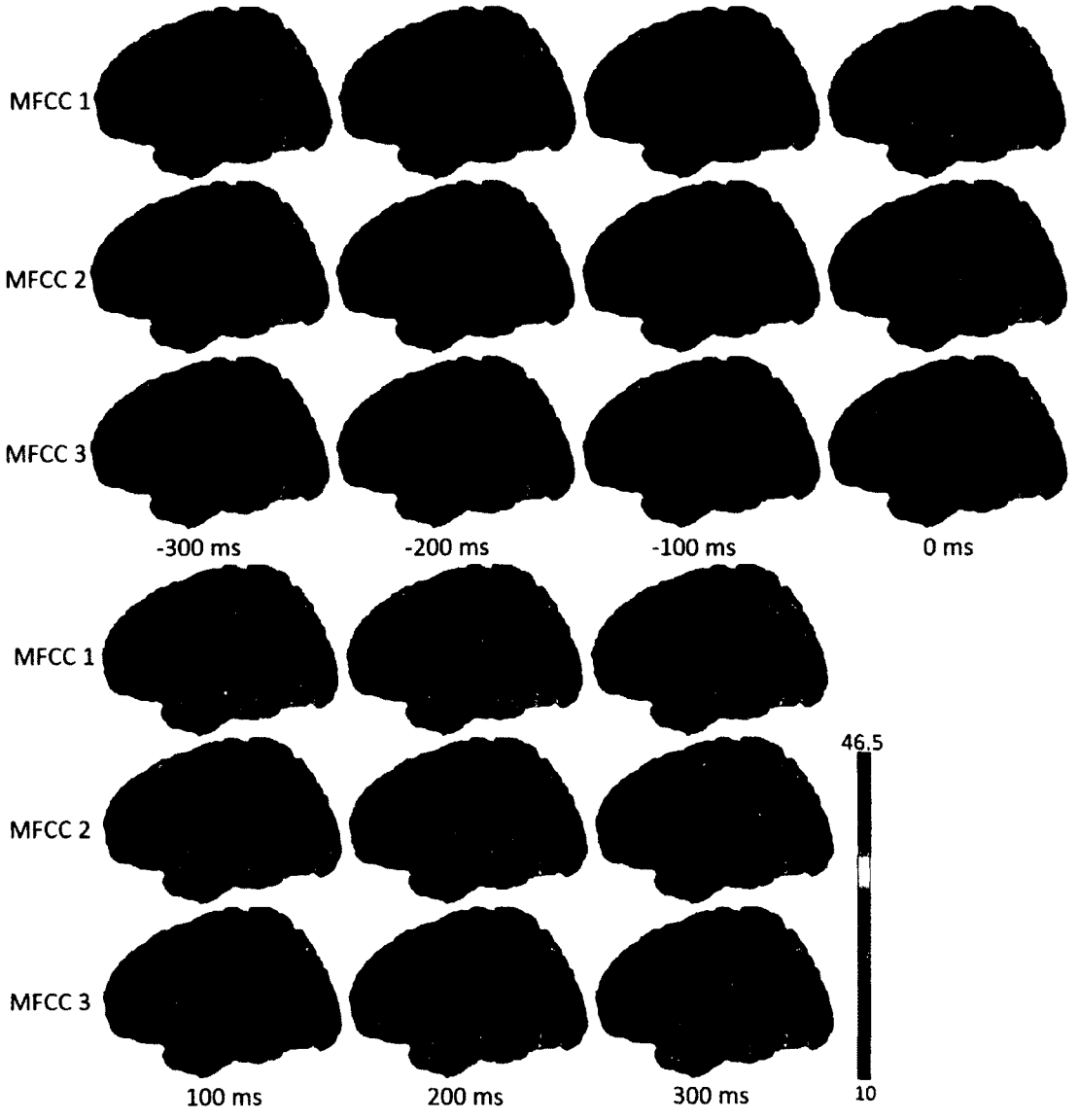


FIG. 31: Spatiotemporal correlations between ECoG high gamma power and the significant MFC coefficients (shown in the three rows respectively) across seven time latencies relative to the onset of speech, with silence periods removed from the analysis.

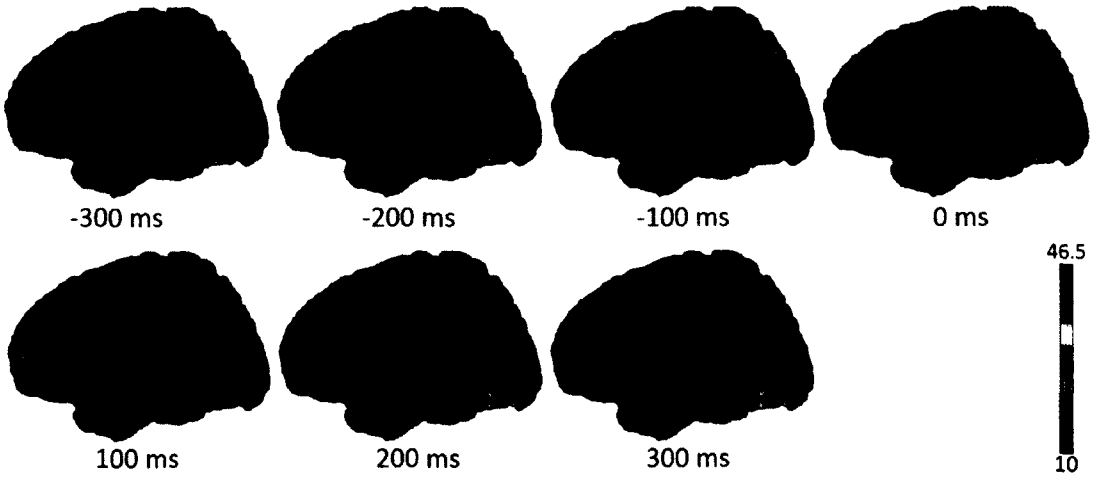


FIG. 32: Spatiotemporal correlations between ECoG high gamma power and the first PLP coefficient across seven time latencies relative to the onset of speech, with silence periods removed from the analysis.

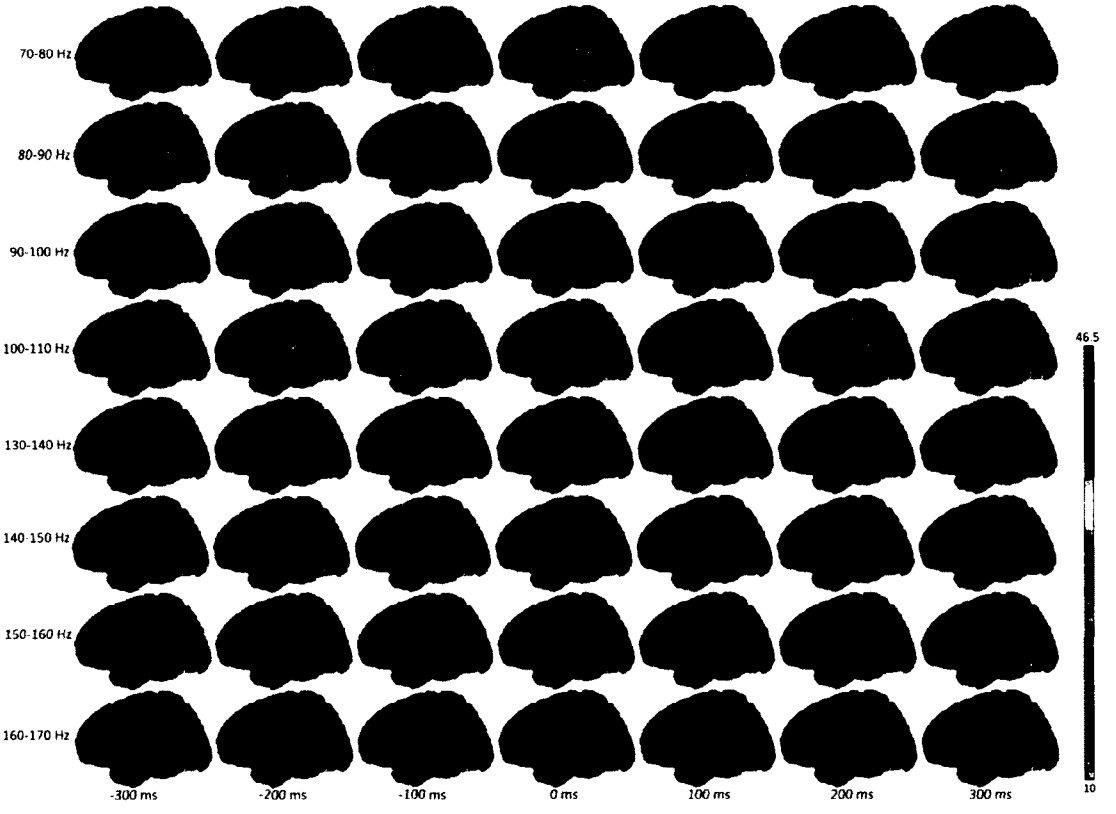


FIG. 33: Spatiotemporal correlations between the eight ECoG high gamma sub-band powers and the speech power, across seven time latencies relative to the onset of speech, with silence periods removed from the analysis.

VITA

Shreya Chakrabarti
Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, VA
Email: schak001@odu.edu

Old Dominion University *August 2015*
Ph.D in Electrical and Computer Engineering GPA: 3.98/4.0
West Bengal University of Technology, India *May 2011*
B.Tech. in Electronics and Communication Engineering GPA: 9.0/10.0

Selected Publications

1. J.S. Brumberg, D.J. Krusienski, A. Gunduz, P. Brunner, **S. Chakrabarti**, A.L. Ritaccio, G. Schalk,"Spatio-temporal Evolution of Cortical Processes Related to Continuous Overt and Covert Speech Production in a Reading Task." (in progress)
2. **S. Chakrabarti**, H. Sandberg, J.S. Brumberg, D.J. Krusienski, "Progress in Speech Decoding from the Electrocorticogram", Biomedical Engineering Letters, Springer, 5(1): 10-21, 2015.
3. **S. Chakrabarti**, GD Johnson, JS Brumberg, G Schalk, and DJ Krusienski. "Exploring the Neural Correlates of Speech and Music in Electrocorticography," International Biomedical Engineering Conference, Korea, 2014.
4. **S. Chakrabarti**, J.S. Brumberg, A. Gunduz, P. Brunner, G. Schalk, D.J. Krusienski, "Modeling the Mel Frequency Cepstral Coefficients of Continuous Speech from Electrocorticographic High-Gamma Activity," 41st Neural Interfaces Conference, Texas, 2014.
5. **S. Chakrabarti**, D.J. Krusienski, G. Schalk, J.S. Brumberg, "Predicting mel-frequency cepstral coefficients from electrocorticographic signals during continuous speech production", Proceedings of the Sixth International Neural Engineering Conference, California, 2013.
6. **S. Chakrabarti**, J.S. Brumberg, A. Gunduz, P. Brunner, G. Schalk, D.J. Krusienski, "Using ECoG Gamma Activity to Model the Mel-Frequency Cepstral Coefficients of Speech", Proceedings of the Fifth International Brain-Computer Interface Meeting: Defining the Future, Pacific Grove, California, 2013.